

Université de Montréal

**A Reliable and Robust Algorithm to Identify Episodes of  
Hospitalizations Using RAMQ Medical Services Claims:  
Methodology Issues & Data Validation**

par

Amir Abbas Tahami Monfared

Programme de Sciences Biomédicales

Faculté de Médecine

Thèse présentée à la Faculté des études supérieures  
en vue de l'obtention du grade de Docteur de Philosophie (Ph.D.)  
en Sciences Biomédicales

avril 2004

© Amir Abbas Tahami Monfared, 2004



W

4

U58

2004

V. 145

## AVIS

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

## NOTICE

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal  
Faculté des études supérieures

Cette thèse intitulée:

A Reliable and Robust Algorithm to Identify Episodes of Hospitalizations  
Using RAMQ Medical Services Claims: Methodology Issues & Data Validation

Présentée par:  
Amir Abbas Tahami Monfared

a été évaluée par un jury composé des personnes suivantes :

Dr Gilles Bleau.....	Président-rapporteur
Dr Jacques LeLorier.....	Directeur de recherche
Dr Lucie Blais.....	Membre du jury
Dr Elham Rahme.....	Examinatrice externe
Dr André-Pierre Contandriopoulos .....	Représentant du doyen de la FES

Thèse acceptée : le 29 septembre 2004.



# ***Abstract***

---

## **Introduction:**

The management of health care system nowadays places major emphasis on reducing cost and resource utilization. On average, the most important characteristics contributing to the total cost in a health care system are in-hospital length of stay (LOS) and number of hospitalizations.

MED-ECHO, which is the hospital discharge summary database in the province of Quebec, contains relevant information regarding hospitalizations, duration, and their related diagnoses (including primary and up to 15 secondary diagnoses). While estimates of in-hospital episodes relying extensively on this database would provide more accurate information, the associated time (about 1 year) for accessing this information currently appears prohibitive in many studies; permission for making a linkage to these data must be granted by the Commission d'accès à l'information (CAI), a procedure that can take several months. Thus, there would be enormous value in finding reliable ways to estimate episodes of hospitalizations in a routine and affordable manner.

Claims data maintained by the Régie de l'assurance Maladie du Québec (RAMQ) have been used in various studies in relatively large populations at limited cost and reasonable time. Most concerns about the validity of research findings, in those cases, were related to the accuracy of diagnostic coding and systematic bias. Although pharmaceutical services information contained in the RAMQ has been subject for validation, the completeness and accuracy of the RAMQ medical services information, particularly concerning inpatient claims remains uncertain. This requires further endeavor since the data concerning hospitalizations are not specifically contained in these data files.

**Objective:**

Time-efficiency in using the RAMQ medical claims data provided the key motivation for this dissertation, which explored the capacity and reliability of claims data, a less costly and albeit less informative data source, to estimate in-hospital LOS and number of hospitalizations (including readmissions) as proxies for hospital morbidity.

The objectives of this study were (1) to develop and validate a reliable and robust algorithm to identify episodes of hospitalizations using the RAMQ physician claims database (training dataset); (2) to examine whether the capacity of the RAMQ claims data for estimating the above-mentioned proxies is being fully recognized; and (3) to assess the applicability and generalizability of the developed algorithm in other clinical conditions, namely acute myocardial infarction and gastrointestinal hemorrhage, using an external validation dataset provided by the RAMQ (validation dataset). All these would increase the potential for acquiring quality information regarding hospitalizations in the RAMQ medical claims database.

**Methodology:**

While the identification of in-hospital codes provides essential information to establish the fact that a patient was certainly admitted to the hospital, we additionally needed an analytical algorithm to identify episodes of hospitalizations. Since it is very likely that a patient does not receive any medical services for a certain period while being hospitalized, to correctly estimate the episodes of care we assumed that there was a service interruption (*gap*) within services provided. Accordingly, several algorithms were built based on a predefined gap between 1 to 15 days. Development of the analytical algorithm to identify in-hospital information with the best possible accuracy required an additional linkage to the MED-ECHO database.

The accuracy of the RAMQ medical claims data concerning inpatient services was assessed using the concepts of diagnostic test evaluation, which was implemented through comparison analyses with a “*gold standard*”, MED-ECHO database. The strengths and limitations of the RAMQ claims data were examined using separate analyses. First, episodes of hospitalizations in a training dataset (a cohort of COPD-diagnosed patients) were analyzed using the developed algorithm. A separate analysis compared agreement between inhospital episodes derived from claims data and the gold standard using an automated deterministic linkage algorithm. This demonstrated whether claims data could be used to identify inhospital morbidity indices. The accuracy of the parameter estimates using claims data were then assessed by comparing the central tendency of the parameters to those observed in the gold standard. The generalizability of the developed algorithm for other clinical conditions (i.e., acute myocardial infarction and upper gastrointestinal bleeding) was assessed in a validation dataset provided by the RAMQ using the concept of external validity.

### **Results:**

Compared with the observed durations and number of hospitalizations in MED-ECHO, the analytical algorithm with the gap of 7 days best defined the episodes of hospitalizations in the RAMQ claims database. This selection was based on the results of the statistical and predictive performance analyses. It was shown that the RAMQ claims data has the capacity for being used as an accurate and reliable source of data for identifying hospitalization episodes. Applying the developed algorithm to an external source of data (provided by the RAMQ) resulted in the close estimates for hospitalization episodes in two other clinical conditions (AMI and UGIB), when compared to those obtained from the MED-ECHO database. The results of the latter analysis simply confirmed the generalizability of the developed algorithm.

**Discussion:**

The present study demonstrated that the RAMQ claims database could be used to accurately measure the LOS and number of hospitalizations for different diseases; it provided reasonable diagnostic agreement with the data collected in the hospital discharge summary database. When estimates of hospitalization episodes from the RAMQ claims data were compared to those obtained from MED-ECHO, the average in-hospital lengths of stay were slightly underestimated in the claims based estimates, whereas the average number of hospitalizations were quite similar. This finding was in total agreement with the assumption that claims data are consistently underestimate the degree of in-hospital morbidities and performance indicators. Furthermore, consistent with prior studies, the findings confirmed that identification of patients with more complicated clinical conditions using ICD-9 codes can be difficult and fraught with unknown effects. Proper design and inclusion criteria should be adapted for patients' selection.

**Conclusion:**

Results of the analyses established the RAMQ claims database as being an excellent source of information for the study of health care resource utilization, particularly hospitalization episodes. The developed algorithm is a valuable tool for identification of in-hospital episodes in the RAMQ claims database. It can easily be integrated in epidemiologic and outcomes research studies. The use of the RAMQ claims database lends itself readily to analyses, providing extremely large data sets at a very little cost. Substantial time and cost savings can be made if information is gathered from the RAMQ.

**Keywords:** Administrative database; physician claims; data quality; analytical algorithm; validation; reliability; hospitalization; length of stay.

# Résumé

---

## Introduction :

De nos jours, la gestion du système de santé met un accent majeur sur la réduction significative des coûts et de l'utilisation des ressources. Par conséquent, pour réduire les coûts, on doit identifier les diverses caractéristiques qui pourraient influencer l'utilisation des ressources. En moyenne, les facteurs les plus importants contribuant aux coûts dans le système de santé sont la durée du séjour (DS) et le nombre d'hospitalisations. Ils sont fréquemment employés comme déterminants de la performance des hôpitaux, de la complexité des cas, de la sévérité de la maladie, de la qualité des soins et des coûts d'utilisation en santé. En outre, examiner les tendances de l'hospitalisation est une priorité importante afin de trouver des moyens appropriés pour l'usage efficace des ressources.

Les données de MED-ECHO (les données récapitulatives de congé d'hôpital) contiennent tous les renseignements pertinents à l'hospitalisation, notamment sa durée, ainsi que le diagnostic primaire et les diagnostics secondaires pour la province de Québec. Alors que les évaluations des épisodes d'hospitalisation se fondant intensivement sur ces données fourniraient des informations plus précises, le temps d'accès à ces informations semble actuellement prohibitif dans la plupart des études; elles sont toujours fournies lorsque la permission est accordée (il est nécessaire d'obtenir la permission de la Commission d'accès à l'information (CAI) pour faire le lien avec une autre source). Ainsi, il y aurait un grand avantage à trouver des moyens fiables d'estimer les épisodes d'hospitalisation de façon courante et accessible.

Les données de réclamations médicales maintenues par le Régie de l'assurance maladie du Québec (RAMQ) ont été employées dans diverses études portant sur des populations relativement grandes, à un coût limité et dans un temps raisonnable. La

plupart des études peuvent être réalisées en croisant les fichiers de base de données de services pharmaceutiques et de services médicaux. Le plus grand souci concernant la validité des résultats de recherche, dans ce cas, s'est concentré sur des questions liées à l'exactitude du codage diagnostique et des erreurs systématiques (biais). Bien que l'information sur les services pharmaceutiques contenue dans les fichiers de la RAMQ ait été soumise à une validation, l'exactitude de l'information sur les services médicaux de la RAMQ, en particulier au sujet des réclamations des personnes hospitalisées, reste incertaine. Ceci exige davantage d'efforts puisque les données se rapportant aux hospitalisations ne sont pas spécifiquement contenues dans ces fichiers de données.

### **Objectifs :**

Le rapport temps-efficacité produit en utilisant les données de la RAMQ pour estimer la DS et les nombres d'hospitalisations (y compris les réadmissions), comme déterminants de la morbidité hospitalière, a fourni la motivation principale à cette thèse qui a exploré la capacité et la fiabilité des données de réclamations, moins coûteuses mais par contre moins instructives.

Les buts de cette étude étaient les suivants : (1) développer et valider un algorithme fiable et robuste pour identifier les épisodes d'hospitalisation en utilisant une base de données des services médicaux de la RAMQ (base de données de formation); (2) examiner si la capacité des données de réclamation de la RAMQ pour estimer les déterminants, mentionnés ci-dessus, est entièrement identifiée; (3) évaluer l'applicabilité de l'algorithme développé pour d'autres conditions cliniques, spécifiquement l'infarctus aigu du myocarde et l'hémorragie gastro-intestinale, en utilisant un ensemble de données de validation externes fourni par le RAMQ (bases de données de validation). L'atteinte de ces buts permettrait d'accroître la qualité de l'information concernant l'estimation des hospitalisations basées sur les données de la RAMQ.

**Méthodologie :**

Bien que l'identification des codes d'établissements fournit des informations essentielles pour établir le fait que le patient a certainement été admis à l'hôpital, nous avons également eu besoin d'un algorithme analytique pour estimer les épisodes d'hospitalisation. Le développement d'un algorithme analytique pour identifier l'information d'hospitalisation avec la meilleure précision a exigé de faire un croisement additionnel avec la base de données de congé d'hôpital de MED-ECHO. Puisqu'il est très probable qu'un patient n'ait aucun service médical inscrit à la RAMQ pendant une certaine période tout en étant hospitalisé, nous avons supposé qu'il y avait une interruption de service (intervalle) parmi les services fournis afin d'estimer correctement les épisodes de soin. Par conséquent, plusieurs algorithmes ont été basés sur un espace prédéfini (intervalle) entre 1 et 15 jours.

L'exactitude des données de réclamations médicales de la RAMQ au sujet des services aux patients hospitalisés a été mesurée en utilisant les concepts de l'évaluation et la validation de tests diagnostiques à l'aide d'analyses de comparaison avec une base de données « gold standard » de MED-ECHO. Les forces et les lacunes des données de réclamations de la RAMQ ont été examinées en utilisant des analyses distinctes. Pour commencer, des épisodes d'hospitalisation dans une base de données de formation (une cohorte de patients atteints de maladie pulmonaire obstructive chronique (MPOC)) ont été analysés en utilisant l'algorithme développé. Ensuite, une nouvelle analyse a comparé l'accord entre les épisodes d'hospitalisation dérivés des données de réclamations (RAMQ) et le « gold standard » (MED-ECHO) à l'aide d'un algorithme automatisé à liaison déterministe. Ceci a démontré que des données de réclamations pourraient être employées pour identifier des index de morbidité hospitalière. L'exactitude des paramètres d'évaluation employant des données de réclamations a par la suite été évaluée en comparant la tendance centrale des paramètres à celle observée dans le « gold standard ». L'applicabilité de l'algorithme développé pour d'autres conditions cliniques (c.-à-d. infarctus aigu du myocarde et

hémorragie gastro-intestinale) a été évaluée, en utilisant le concept de validité externe, dans un ensemble de données de validation fourni par le RAMQ.

### **Résultats :**

Comparativement aux durées et aux nombres d'hospitalisations observés dans MED-ECHO, l'algorithme analytique avec un intervalle de 7 jours est celui qui a le mieux défini les épisodes d'hospitalisation dans la base de données de réclamations de la RAMQ. Le choix de 7 jours a été basé sur les résultats des analyses de performances statistiques et prédictives. Il a été démontré que les données de réclamation de la RAMQ peuvent être utilisées comme source précise et fiable de données pour identifier les épisodes d'hospitalisation. L'application de l'algorithme analytique avec une nouvelle source externe de données (fournie par la RAMQ) a résulté, en faisant la comparaison avec MED-ECHO, à des estimations similaires quant aux épisodes d'hospitalisations pour deux conditions cliniques étudiées (infarctus aigu du myocarde et hémorragie gastro-intestinale). Les résultats de cette dernière analyse ont simplement confirmé la généralisabilité de l'algorithme développé.

### **Discussion :**

La présente étude a démontré que la base de données de services médicaux de la RAMQ peut efficacement être employée pour mesurer la DS et les nombres d'hospitalisations pour différentes maladies; ces données sont raisonnablement en accord avec les données recueillies dans la base de données sommaires des congés d'hôpital (MED-ECHO). Lorsque les évaluations des épisodes d'hospitalisation des données de réclamation de la RAMQ ont été comparées à celles obtenues de MED-ECHO, les durées moyennes de séjour ont été légèrement sous-estimées dans les évaluations basées sur la RAMQ alors que le nombre moyen d'hospitalisations était



tout à fait semblable. Ces résultats sont totalement en accord avec la prétention que les données de réclamations sous-estiment régulièrement le degré de morbidité hospitalière et les index de performance. De plus, en accord avec les études antérieures, les résultats ont confirmé que l'identification des patients dans des conditions cliniques plus compliquées, en employant les codes ICD-9, peut être difficile et comporter des conséquences inconnues. Des critères de conception et d'inclusion appropriés devraient être adaptés pour la sélection des patients.

### **Conclusion**

Les résultats des analyses ont établi que la base de données de réclamations de la RAMQ est une excellente source d'information pour l'étude de l'utilisation des ressources en santé, en particulier pour les épisodes d'hospitalisation. L'algorithme développé est un outil valable pour l'identification des épisodes d'hospitalisation dans la base de données de réclamations de la RAMQ. Il peut facilement être intégré dans des études épidémiologiques et économiques. L'utilisation de la base de données de la RAMQ se prête aisément aux analyses, fournissant des données vastes à un coût raisonnable. Des économies de temps et de coûts substantiels peuvent être ainsi faites par rapport à l'utilisation de la banque de données MED-ECHO.

**Mots clés :** Base de données administratives; réclamations médicales; qualité des données; algorithme analytique; validation; fiabilité; hospitalisation; durée du séjour.

# *Acknowledgement*

---

The completion of this thesis would have not been possible without the generous contribution, support, and guidance of many individuals to whom I would like to extend my gratitude and appreciation. First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Jacques LeLorier, whose enormous clinical experience, scientific enthusiasm, and continuous support and encouragement have been invaluable. I truly appreciate his positive and insightful guidance, inspiring leadership, “open door” policy, as well as his vast knowledge in the field of clinical pharmacology. I am also thankful for his high standards, communication skills, and patience. They are trademarks that I constantly strive to attain for myself.

I would also like to acknowledge the contribution of my committee members for their generosity, time, and energy. These people generously undertook to read the manuscript entirely and to give me the benefit of their expert criticisms and suggestions. Each of them deserves major credit for the quality of the content of this dissertation, without bearing any responsibility for the shortcomings it contains.

A great appreciation is also due to Dr. Lucie Blais (University of Montreal, Faculty of Pharmacy) and Dr. Elham Rahme (McGill University, McGill University Health Center) who not only provided the data for this study but also accommodated insights throughout my research. For their availability, guidance, and insightful comments I am most grateful.

Special thanks are also due to Marc Dorais, my friend and colleague, who provided me a priceless tutorial in SAS (Statistical Analysis Software) programming language. The expertise he provided and his willingness to teach were integral to this

study. I sincerely appreciate all the time and effort from him. His everlasting enthusiasm for statistics has been a continuous motivation for me!

I would also like to thank my friends and colleagues, Delphine Gueylard-Chenevier, Caroline Quach, Héloïse Cardinal, and Hani Tamim for stimulating discussion and pleasant atmosphere that have been of great value to me while struggling with this study. Caroline and Hani were also good enough to read segments of early versions of the manuscript. They gave me the running start I needed in order to transform rough drafts into a finished material.

The following people also made significant contributions to my work and warrant my deepest appreciation: Mrs. Anita Massicotte for her outstanding organizational and interpersonal skills, patience, and secretarial assistance; Odile Sheehy, Frédéric Lavoie, and Claudie Charbonneau for stimulating discussions, support, and pleasant work environment, particularly during our group meetings.

All these individuals from a wide variety of disciplines have constantly provided invaluable assistance combined with criticisms, creative suggestions, mentoring, and encouragement during my matriculation. My gratitude to them is boundless. To each of you, "Thank you".

I cannot begin to express my gratitude to my family and spouse for their continued support and encouragement in my academic and professional choices. I hereby wish to express my heartfelt thanks to my parents; they have followed my academic career and have always been encouraging, offering unconditional love and support. Last but not least, my deepest and most loving thanks are due to my wife, Sanaz, for her encouragement and inspiration throughout the entire course of my studies, everlasting support, and unconditional love. Certainly, this dissertation could not have been completed without her.

*To my parents,  
who made me who I am*

# Contents

---

<b>Abstract .....</b>	<b>i</b>
<b>Résumé .....</b>	<b>v</b>
<b>Acknowledgement .....</b>	<b>x</b>
<b>Contents.....</b>	<b>xiii</b>
<b>List of Tables.....</b>	<b>xx</b>
<b>List of Figures .....</b>	<b>xxi</b>
<b>Symbols &amp; Abbreviations .....</b>	<b>xxiii</b>
<b>CHAPTER 1: Introduction .....</b>	<b>1</b>
<b>CHAPTER 2: Review of Literature.....</b>	<b>9</b>
2.1 CHRONIC OBSTRUCTIVE PULMONARY DISEASE .....	9
2.1.1 Definition.....	9
2.1.2 Etiology.....	11
2.1.3 Epidemiology.....	13
2.1.3.1 Prevalence .....	13
2.1.3.2 Mortality.....	13
2.1.3.3 Morbidity.....	13
2.1.4 Burden of Chronic Obstructive Pulmonary Disease .....	14
2.2 CLINICAL INDICATORS IN OUTCOMES RESEARCH .....	18
2.2.1 Introduction .....	18
2.2.2 Background.....	20
2.2.2.1 Definition: Clinical Indicator .....	20

2.2.2.2 Process or outcome indicators .....	21
2.2.2.2.1 Outcomes of care .....	22
2.2.2.2.2 Process of care .....	23
2.2.2.3 Domains of quality .....	24
2.2.2.4 Rate-based or sentinel event indicator.....	24
2.2.3 <i>Common Themes</i> .....	25
2.2.3.1 Evidence-based process measures.....	25
2.2.3.2 Generic outcome measures.....	26
2.2.3.2.1 Mortality .....	26
Clinical significance .....	29
2.2.3.2.2 Inhospital Length of Stay (LOS).....	30
Clinical significance .....	33
2.2.3.2.3 Readmission to hospital .....	34
Clinical significance .....	37
2.3 DATA COLLECTION AND SAMPLING.....	39
2.3.1 <i>Administrative Databases</i> .....	40
2.3.2 <i>Canadian Administrative databases</i> .....	45
2.3.3 <i>Accuracy and Sources</i> .....	46
2.3.4 <i>Substantive Databases</i> .....	49
2.3.4.1 Populations and Events .....	49
2.3.4.2 Hospital Discharge Data.....	50
2.3.4.3 Pharmaceutical Data.....	56
2.3.4.4 Physician Claims .....	58
2.3.4.5 External Criteria .....	60
2.3.4.6 Sensitivity Testing.....	61
2.3.4.7 Summary of Databases by Province.....	62
2.4 ACCURACY ASSESSMENT OF ADMINISTRATIVE DATABASES .....	63
2.4.1 <i>Electronic Database Comparison Method</i> .....	64
2.4.2 <i>Chart Audit Method</i> .....	66
2.4.2.1 Demographic Data.....	66

2.4.2.2 Diagnostic Data .....	68
2.4.2.3 Procedural Data .....	70
2.4.3 <i>Expert Clinical Criteria Method</i> .....	70
2.5 ANOMALIES IN ADMINISTRATIVE DATABASES .....	73
2.5.1 <i>Error</i> .....	74
2.5.1.1 Acquisition Errors .....	74
2.5.1.1.1 Data Entry .....	74
2.5.1.1.2 Invalid Value.....	75
2.5.1.1.3 Misrepresentation.....	76
2.5.1.1.4 Consistency .....	76
2.5.1.2 Transformation Errors .....	77
2.5.1.2.1 Transformation Algorithms.....	77
2.5.1.3 Analysis Errors .....	78
2.5.1.3.1 Wrong Assumptions.....	78
2.5.2 <i>Missing Data</i> .....	78
2.5.2.1 Acquisition Missing Data.....	78
2.5.2.1.1 Not Applicable .....	79
2.5.2.1.2 Unknown.....	79
2.5.2.1.3 Not Measured.....	79
2.5.2.1.4 Not Noteworthy .....	79
2.5.2.1.5 Not Necessary .....	79
2.5.2.1.6 Not Recorded .....	80
2.5.2.1.7 Mis-recorded .....	80
2.5.2.1.8 Lost Data .....	80
2.5.2.1.9 Normal Values .....	81
2.5.2.2 Transformation: Missing Data.....	81
2.5.2.2.1 Exclusion of Rows and /or Records .....	81
2.5.2.2.2 Exclusion of Fields .....	81
2.5.2.2.3 Exclusion of Values .....	82
2.5.2.3 Analysis of Missing Data .....	82
2.5.2.3.1 Eliminating Rows/Records .....	83

2.5.2.3.2 Eliminating Fields .....	83
2.5.2.3.3 Eliminating Values.....	83
2.5.3 Variation.....	84
2.5.3.1 Acquisition Variation .....	84
2.5.3.1.1 Structural.....	85
2.5.3.1.2 Definition .....	86
2.5.3.1.3 Method / Measurement .....	86
2.5.3.1.4 Constraints / Rules .....	86
2.5.3.1.5 Temporal / Historical .....	87
2.5.3.1.6 Record Keeping Practices .....	87
2.5.3.1.7 Purpose.....	88
2.5.3.2 Transformation Variation .....	88
2.5.3.2.1 How Data Are Transformed.....	88
2.5.3.3 Analysis Variation.....	89
2.5.3.3.1 How Data Are Used.....	89
2.5.3.3.2 Assumptions.....	89
PROCESS.....	90
Acquisition.....	90
Transformation (zero to many) .....	90
Analysis .....	90
(zero to many).....	90
Potential Causes of Anomalies.....	90
Errors .....	90
Missing .....	90
2.5.4 Assessing a Condition in Administrative Databases Using a Gold Standard Test.....	91
<b>CHAPTER 3: Objective &amp; Hypothesis.....</b>	<b>94</b>
3.1 PROBLEM.....	94
3.2 HYPOTHESIS AND OBJECTIVE.....	95
<b>CHAPTER 4: Methodology.....</b>	<b>98</b>
4.1 INTRODUCTION.....	98



4.2 DATA SOURCES .....	101
4.2.1 <i>MED-ECHO</i> .....	102
4.2.2 <i>RAMQ</i> .....	104
4.2.2.1 Medicare Registrant Database (Enrollment) .....	106
4.2.2.2 Physician Claims Database .....	107
4.2.2.3 Physician Demographic Database .....	109
4.2.2.4 Prescription Claims Database .....	109
4.3 DEFINITIONS .....	111
4.3.1 <i>Available Information on Claims Data</i> .....	111
4.3.2 <i>Accuracy and Agreement</i> .....	112
4.4 THE LOGIC OF COMPARISON .....	114
4.5 ALGORITHM .....	117
4.5.1 <i>Introduction</i> .....	117
4.5.2 <i>Study Sample</i> .....	118
4.5.3 <i>Development</i> .....	119
4.5.4 <i>Statistical Analysis</i> .....	128
4.5.4.1 Descriptive Statistics .....	129
4.5.4.2 Predictive Performance .....	130
4.5.4.3 Measures and Levels of Agreement .....	133
4.5.4.3.1 Efficiency (Overall Agreement) .....	135
4.5.4.3.2 Sensitivity .....	135
4.5.4.3.3 Specificity .....	136
4.5.4.3.4 Predictive Values .....	138
4.5.4.3.5 Likelihood Ratio .....	138
4.5.4.3.6 Kappa .....	139
4.5.4.3.7 Kullback-Leibler Distance (KLD) .....	141
4.5.4.3.8 Levels of Agreement .....	142
4.5.5 <i>Application</i> .....	144
4.5.5.1 Study Sample .....	145
4.5.5.2 Analyses .....	146
4.5.5.2.1 Primary .....	146

4.5.5.2.2 Secondary .....	147
<b>CHAPTER 5: Result .....</b>	<b>153</b>
5.1 DEVELOPED ALGORITHM (COPD) .....	153
5.1.1 Central Tendency .....	154
5.1.2 Predictive Performance .....	160
5.1.3 Analysis of Agreement .....	171
5.2 APPLICATION .....	173
5.2.1 Population Characteristics .....	174
5.2.2 Central Tendency .....	175
5.2.3 Hospitalizations for Specific Indications .....	183
5.2.3.1 Acute Myocardial Infarction (AMI) .....	183
5.2.3.1.1 Diagnosis Agreement .....	184
5.2.3.1.2 Central Tendency .....	185
5.2.3.2 Upper Gastrointestinal Bleeding (UGIB) .....	189
5.2.3.2.1 Diagnosis Agreement .....	189
5.2.3.2.2 Central Tendency .....	194
<b>CHAPTER 6: DISCUSSION .....</b>	<b>198</b>
6.1 HOSPITALIZATION EPISODES .....	200
6.1.1 The Optimal Gap .....	201
6.1.2 Appropriateness of Central Tendency .....	205
6.1.2.1 Typical Measure .....	207
6.1.2.2 Outliers .....	208
6.1.3 Diagnosis Agreement .....	210
6.1.3.1 COPD Hospitalizations .....	211
6.1.3.2 AMI Hospitalizations .....	213
6.1.3.3 UGIB Hospitalizations .....	214
6.2 LIMITATIONS .....	218
6.3 IMPLICATIONS .....	221
<b>CHAPTER 7: Conclusion .....</b>	<b>224</b>

<b>Reference.....</b>	<b>227</b>
<b>Appendix 1 .....</b>	<b>254</b>
<b>Appendix 2 .....</b>	<b>255</b>
<b>Appendix 3 .....</b>	<b>256</b>
<b>Appendix 4 .....</b>	<b>257</b>

# *List of Tables*

---

<b>Table 2.1</b> – Assessment of population coverage.....	49
<b>Table 2.2</b> – Comparison of hospital discharge summary with an additional source.....	51
<b>Table 2.3</b> – Summary of Canadian reabstraction studies of hospital records.....	52
<b>Table 2.4</b> – Summary of Canadian reabstraction studies (specific).....	53
<b>Table 2.5</b> – Studies on the accuracy of physician claims in Canada.....	59
<b>Table 2.6</b> – Summary of abnormalities found in administrative databases.....	90
<b>Table 4.1</b> – List of standard data elements in the RAMQ medical database.....	112
<b>Table 5.1</b> – Estimated LOS in the RAMQ medical claims database.....	156
<b>Table 5.2</b> – Estimated number of hospitalizations in the RAMQ medical claims database..	157
<b>Table 5.3</b> – Absolute predictive performance of selected algorithms (gaps).....	161
<b>Table 5.4</b> – Estimated population parameters in a cohort of COPD patients using the RAMQ database, compared to the Gold Standard.....	170
<b>Table 5.5</b> – Estimated population parameters in the RAMQ database compared to the values observed in the Gold Standard.....	176
<b>Table 5.6</b> – Estimated population parameters for AMI in the RAMQ database compared to the values observed in MED-ECHO.....	186
<b>Table 5.7</b> – Selected agreement statistics comparing different models for UGIB.....	191
<b>Table 5.8</b> – Reliability statistics comparing different models for UGIB.....	192
<b>Table 5.9</b> – Estimated population parameters for UGIB in the RAMQ database compared to the parameters observed in MED-ECHO.....	195

# *List of Figures*

---

<b>Figure 4.1</b> – Four stages in a database validation study using linkage method.....	100
<b>Figure 4.2</b> – Algorithm to estimate episodes of hospitalizations .....	121
<b>Figure 4.3</b> – Schematic overview of analysis of agreement in claims data.....	134
<b>Figure 5.1</b> –Age distribution in the COPD cohort by gender.....	154
<b>Figure 5.2</b> –Estimated length of stay (LOS) in RAMQ using different gaps.....	158
<b>Figure 5.3</b> –Estimated N° of hospitalizations in RAMQ using different gaps.....	159
<b>Figure 5.4</b> –Distribution of predictive performance (% PE) for selected algorithms.....	162
<b>Figure 5.5</b> – Residual error estimates of selected algorithms (LOS) .....	163
<b>Figure 5.6</b> – Residual error estimates of selected algorithms (N° of Hospitalizations) .....	164
<b>Figure 5.7</b> – Frequency distribution of gaps in the RAMQ medical claims database.....	166
<b>Figure 5.8</b> – Frequency distribution of hospitalization intervals in MED-ECHO.....	167
<b>Figure 5.9</b> – Histogram of average LOS and its <i>pdf</i> .....	168
<b>Figure 5.10</b> – Histogram of N° of hospitalizations and its <i>pdf</i> .....	168
<b>Figure 5.11</b> – Comparison of the estimated LOS and N° of hospitalizations to the gold standard (MED-ECHO). As shown in these graphs, LOS and number of hospitalizations both follow a lognormal distribution.....	169
<b>Figure 5.12</b> – Estimated LOS and N° of hospitalizations compared to MED-ECHO database (Gold Standard) .....	178
<b>Figure 5.13</b> – Histogram and the <i>pdf</i> of LOS in MED-ECHO and RAMQ databases.....	179
<b>Figure 5.14</b> – Histogram and the <i>pdf</i> of N° of hospitalizations in both databases.....	179

<b>Figure 5.15</b> – Comparison of the estimated LOS and N° of hospitalizations with the gold standard, both <i>pdfs</i> follow lognormal distribution.....	180
<b>Figure 5.16</b> – Frequency distribution of gaps in the RAMQ medical claims database.....	182
<b>Figure 5.17</b> – Frequency distribution of hospitalization intervals in MED-ECHO.....	182
<b>Figure 5.18</b> – Histogram and <i>pdf</i> distribution of LOS in MED-ECHO and RAMQ.....	187
<b>Figure 5.19</b> – Comparison of the estimated LOS inth RAMQ with MED-ECHO as the gold standard, LOS follows a lognormal distribution.....	188
<b>Figure 5.20</b> – Histogram and the pdf distribution of LOS in MED-ECHO and RAMQ for models M2 (Most sensitive) and M4 (most specific).....	196
<b>Figure 5.21</b> – Comparison of the estimated LOS in the RAMQ with MED-ECHO as the gold standard; distribution of LOS can be approximated using a lognormal distribution.....	197
<b>Figure 6.1</b> – Comparison of percent prediction errors (%PE) using different gaps.....	202

## ***Symbols & Abbreviations***

---

<b><math>\kappa</math></b>	Cohen's kappa
<b><math>\Sigma</math></b>	Summation symbol
<b><math>\sqrt{\phantom{x}}</math></b>	Square root
<b><math>  </math></b>	Absolute value
<b><math>\times</math></b>	Operation for multiplication group
<b>% APE</b>	Absolute prediction error percentage
<b>% PE</b>	Prediction error percentage
<b>CI</b>	Confidence interval
<b>CV</b>	Coefficient of variation
<b>FN</b>	False negative
<b>FP</b>	False positive
<b>KL distance</b>	Kullback-Leibler distance
<b>LR+</b>	Likelihood ratio of positive test
<b>LR-</b>	Likelihood ratio of negative test
<b>MAE</b>	Mean absolute error
<b>ME</b>	Mean error
<b>PABAK</b>	Prevalence and bias adjusted kappa
<b>pdf</b>	probability density function

<b>P<sub>e</sub></b>	Expected (Chance) agreement
<b>P<sub>o</sub></b>	Observed agreement
<b>P<sub>in</sub></b>	Rule-in probability
<b>P<sub>out</sub></b>	Rule-out probability
<b>PVP</b>	Predictive value of positive test
<b>PVN, NPV</b>	Predictive value of negative test
<b>RMSE</b>	Root mean squared error
<b>SD, <math>\sigma</math></b>	Standard deviation
<b>Se</b>	Sensitivity
<b>Sp</b>	Specificity
<b>TN</b>	True negative
<b>TP</b>	True positive



# ***Introduction***

---

# **1**

Today's research increasingly focuses on the observation of outcomes to assess the impact of medical procedures and interventions. *Outcomes Research* seeks to understand, describe, interpret, and predict the impact of various practices, especially (but not exclusively) interventions on «final» end results that matter to decision makers: patients, providers, private payers, government agencies, or society at large.

Measuring outcomes—and applying the results to continually improve care—is an on-going challenge. Assessing the outcome of care is particularly important for chronic diseases, such as chronic obstructive pulmonary disease (COPD), because of their increasing burden upon the health care system. COPD is a chronic progressive disorder and in spite of the large number of patients undergoing active medical interventions at any given time and various guidelines for the treatment and management of COPD [Briggs DD, 2000], there is a lack of consensus regarding the detailed components of effective Outcomes research. It is not known, for example, whether management plans are effective in reducing health care costs (e.g.; by decreasing the length of inhospital stay) or improving the outcomes of exacerbations (e.g.; by reducing the number of inhospital visits). A patient with COPD is frequently encountered by the physician, and may further require inhospital admission for an

acute exacerbation episode. Therefore, for an outcome research study in COPD, the collection of data must cover the primary care setting including in-hospital records, which should strongly depend on solid information base—*valid data*. In-hospital information is inevitably important since it can be used for *Effective Outcomes Research* by which one can readily assess health care utilization, evaluate costs, and last but not least, examine in-hospital morbidity and disease complexity.

Effective Outcomes Research becomes even more important, when we realize that a more broadly defined set of outcomes than short-term morbidity and mortality indexes are necessary to develop efficient ways of providing services to patients. For instance, Outcomes data including hospital discharge are needed to fully evaluate some effects of inpatient interventions and services. Single indicators, indeed, in isolation often do not provide the answer, particularly when one considers a high variation in surrogate measures in different clinical settings. *Accordingly, measuring Outcomes indicators, particularly in the management of certain diseases, with exactitude is essential and requires modern technologies and analytical instruments.* The traditional approach to *Outcomes Research* is often based on analysis of secondary (*administrative claims*) data, through data collected from patient surveys, or medical chart reviews [Wray NP, 1995].

Accordingly, the development of complex health care systems in several countries has contributed to the creation of large health databases, such as those maintained by the provincial health insurance database in Quebec (Régie de

l'assurance Maladie du Québec—RAMQ), for administrative purposes. Although they were not originally intended for research use, and despite their limitations [Bright RA, 1989; Potvin L and Champagne F, 1986; Demlo LK and Campbell PM, 1981], administrative databases are frequently used in analytical epidemiologic studies. Indeed, they can provide an increasingly accessible and widely used source of data for health care research and technology assessment [Roos LL, 1989; Connell FA, 1987]. Because of the large size and nearly universal coverage of defined populations in these databases, they offer the opportunity for population-based research, with large sample sizes and possible longitudinal follow-up. Moreover, administrative databases benefit from other advantages namely: non-intrusive data collection, low cost, and no reliance on individual recall [Roos LL and Sharp SM, 1988; Roos LLJr, 1987].

Medicare files provided by the RAMQ have been used in various studies (i) to investigate disease incidence, adverse drug reaction, and systematic bias [Garbe E, 1998; McLean M, 1994; Beland F, 1989], (ii) to document variations in treatment patterns [Sheehy O and LeLorier J, 2000], (iii) or regional variation in the management of certain diseases [Rodrigues EJ, 2002], (iv) to evaluate the outcomes of specific medical treatments [Garbe E, 1997], (v) to examine diagnostic and procedure coding validation [Tamblyn R, 2000], (vi) to investigate the utilization patterns and effects of drugs [Blais L, 2003; Rahme E, 2002; Blais L, 2001; LeLorier J, 1995], and (vii) to assess the effect of different factors on health care utilization [Blais R and Maiga A, 1999; Demers M, 1995] in relatively large populations at limited cost and

reasonable time. Most concern about the validity of research findings, in those cases, has focused on issues related to the accuracy of diagnostic coding and systematic bias. In a number of studies, drug prescription claims information included in the RAMQ has also been linked to other health care databases, and has been subject for validation [Tamblyn R, 1995]. However, the completeness and accuracy of the RAMQ medical services information, particularly concerning inpatient (hospitalization) records remains uncertain. This requires further endeavor since the data concerning hospitalizations are not specifically contained in these data files and must be deduced from the medical claims designated to in-hospital visits by the physician. Moreover, a readmission occurring within a few days after the discharge complicates the process.

Although the identification of in-hospital codes provides essential information to establish the fact that the patient was certainly admitted to the hospital, we additionally need a computer algorithm to estimate the duration of hospitalizations. Development of an algorithm to identify in-hospital information with the best possible accuracy requires an additional linkage to the Quebec hospital discharge database (known as MED-ECHO: Maintenance et Exploitation des Données pour l'Étude de la Clientèle Hospitalière) in which all the relevant records regarding hospitalizations, duration, and their related diagnoses are stored. Although this linkage is not technically difficult, it is time-consuming and usually imposes delays in research projects (inefficiency). Investigators, including those at the Epidemiology and Economics Research Unit at Centre Hospitalier de l'Université de Montréal (CHUM),

have used this research methodology to report the health outcomes in a number of publications using administrative databases and record linkage.

Using administrative databases, however, makes one to contend with the problem of overcoming bias from the many confounding factors, such as the question of data validity. Validation of data in any epidemiologic study is very important; and the RAMQ medical services database is not an exception. Fortunately, the structure of RAMQ database permits examination of some aspects of the completeness and accuracy of its data through external validation. The validation of RAMQ database, concerning hospitalization of a target population, can be implemented through comparison analyses with a “*gold standard*”, i.e., MED-ECHO database. Since it is unlikely that the two data streams would agree and that both would be incorrect, the integration of these two independent data streams could also be used to investigate the **outcomes** of care.

Outcome measures are often seen as the ‘Holy Grail’ for quality of care measurement [Berwick DM, 1989]. Generic markers of patient outcomes are commonly used as surrogate measures of quality-of-care because of the relative ease of obtaining this information. Examples of these quality-of-care parameters are mortality and morbidity. Morbidity is often described by the inhospital length of stay (LOS) or hospital readmission; which both can be estimated by the RAMQ medical services database. Moreover, the RAMQ database can provide the opportunity for

comparative analyses through the use of summary statistics to identify trends over time and between institutions.

The **length of stay** (LOS) is an important variable in the analysis of hospital performance; it is frequently used as an indicator of *efficiency*, and also used as a proxy to other aspects of care, such as *case complexity*, *severity of illness* and *quality of care*. Usually, it is tacitly assumed that the actual length of stay is the number of days it takes for the hospital to provide proper care for the patient, including time for diagnostic examination, treatment, recovery, and whatever is needed.

In the changing healthcare environment, efficient utilization of inpatient resources is an essential task. The length of stay can be used as a proxy for hospital costs and efficiency. Reducing the length of hospital stay is a fundamental element of modern health cost containment (or policy). Managing of health care resources has become a daunting task and the ability to gather and meaningfully evaluate the pertinent administrative data is essential for operational success. Relevant objective data must be collected, analyzed, and interpreted appropriately. The information thus obtained can be used to estimate the average hospitalization length of stay and consequently the relative cost of health care services provided to the target population.

Furthermore, clinical curiosity and the high economic cost of patient **readmission** have stimulated the interest in hospital readmissions over the past decades. Research studies, for this purpose, are extremely facilitated by the ease and

relative small costs associated with extracting the data from existing administrative systems. These factors and the apparent face validity of hospital readmissions as a quality-of-care measure promulgated the use of readmission rates as a clinical indicator [Ibrahim JE, 1999]. Further studies can then be implemented in order to examine the appropriateness of the hospital readmissions.

Evaluating hospital length of stay and readmission is a challenging task, particularly since some patients could have rather extended stays. This dissertation explores this issue in an effort to investigate the quality of RAMQ administrative data. In so doing, we developed an analytical algorithm to evaluate the length of stay and number of hospitalizations in a group of patients based on the gold standard database, MED-ECHO. Although acquiring information on resource utilization is much easier and faster when data are gathered using administrative databases (compared to prospectively data collection); validation of the RAMQ medical claims database is essential since information recorded in this database are relatively more accessible compared to the MED-ECHO database that is often time-consuming. Substantial time and cost savings can be made if information could instead be collected from RAMQ, although this approach has possible shortcomings as well as advantages.

The consistency between inpatient resource utilization found in the RAMQ administrative data and records from MED-ECHO database could be based on choice of selected diagnostic codes. The main goal of this project is to develop and to validate a reliable and robust algorithm to determine in-hospital episodes of

hospitalizations, including length of stay and readmission using the RAMQ medical services database. This can be made through an automated deterministic linkage model with the hospital discharge summary, MED-ECHO database; the sensitivity, specificity, and positive and negative predictive values of the RAMQ database can then be estimated, as well as measures of reliability such as kappa ( $\kappa$ ) statistics.



# ***Review of Literature*** **2**

---

## **2.1 Chronic Obstructive Pulmonary Disease**

### **2.1.1 Definition**

Chronic obstructive pulmonary disease (COPD) is a progressive lung disease characterized primarily by airway obstruction and decreased airflow [Ferguson GT and Cherniack RM, 1993]. Its symptoms do not usually become apparent until the age of 55. The changes to the lung, however, actually begin many years earlier. Chronic bronchitis and emphysema are two distinct components of COPD, but may occur simultaneously in the same individual.

Chronic bronchitis is defined as “the presence of a chronic cough for 3 months in each of two successive years in a patient in whom other causes of chronic cough have been excluded” [American Thoracic Society, 1995]. Chronic bronchitis is characterized by chronic inflammation and edema of the peripheral airways, excessive mucus production and accumulation, bronchospasm, bronchial airway obstruction, and hyperinflation of the alveoli distal to the obstructed airways.

Emphysema is defined as “the abnormal permanent enlargement of the airspaces distal to the terminal bronchioles, accompanied by the destruction of their walls and without obvious fibrosis” [American Thoracic Society, 1995]. It is

characterized by alveolar deterioration and hyperinflation, destruction of pulmonary capillaries, weakened respiratory bronchioles, and air trapping.

COPD develops slowly over a period of years and is irreversible, often resulting in death. As the disease progresses, individuals with COPD experience reduced quality of life due to limited activity levels. The primary physical manifestations that result from pathophysiologic changes in COPD include dyspnea (shortness of breath), persistent cough, excessive sputum production, fatigue, decreased exercise tolerance, hypoxemia, and deconditioning [Gerald LB, 2001; Schulman LL, 2000].

COPD is a progressive disease that worsens in severity with time, and is characterized by recurrent “exacerbations” of varying intensity. An acute COPD exacerbation is defined as any combination of worsening dyspnea, increase in sputum production, and/or increase in sputum purulence [Snow V, 2001]. In Canada it is estimated that individuals with COPD experience anywhere from 1-4 acute exacerbations of the disease each year [Canadian Respiratory Review Panel, 1998].

Thus, exacerbations are detrimental to a COPD patient’s quality of life and may result in either temporary or permanent disability, increased emergency room visits and hospital admissions, respiratory failure, or even death. An important goal of COPD management therefore is to decrease the number and the severity of

exacerbations experienced by COPD patients, through comprehensive patient education, early aggressive medical management, and proper follow-up care.

### **2.1.2 Etiology**

The etiology of COPD can be attributed to smoking in 80-90% of cases. COPD can also be induced by exposure to second-hand smoke and/or environmental pollutants, a history of recurrent respiratory infections in childhood, as well as a genetic disorder in 1-5% of cases [Lacasse Y, 1999]. Approximately 15% of all individuals who smoke will develop COPD [Sherk PA and Grossman RF, 2000] and they are 10 times more likely to die of COPD than nonsmoker. Smoking cessation has been associated with a decreasing frequency of recurrent respiratory symptoms and infections in former smokers, as compared to those individuals who continued to smoke [Health Canada, 2001].

Smoking is not the only etiology that is related to the development of COPD. Hereditary disorders also play a role in the onset of COPD early in life. The most common genetic cause of COPD is related to an inherited deficiency in alpha<sub>1</sub>-antitrypsin (ATT), a protein that is normally produced by the liver that plays a role in the inhibition of several proteases, including neutrophil elastase [Sandford AJ and Pare PD, 2000].

The remaining causes of COPD include prolonged exposure to second-hand smoke and/or environmental pollutants and recurrent respiratory infections during childhood. Second-hand smoke can be just as harmful to nonsmokers. It contains 200 poisonous chemicals that can cause serious health problems including respiratory infections, COPD exacerbations, and asthma. Other environmental pollutants such as carbon monoxide, sulfur dioxide, nitrogen oxide, the ozone, and suspended particulates may initiate the inflammatory process and contribute to the development of COPD [Lohr KN, 1992].

Recurrent respiratory infections during childhood have also been suggested as a cause of COPD. Early recurrent infections in childhood can stunt the growth of lung tissue and result in decreased forced expiratory volume in 1 second ( $FEV_1$ ) and forced vital capacity (FVC) in adulthood [Shaheen SO, 1998]. Permanent damage and fibrotic scarring of the airways may occur as a result of the excessive inflammatory response (airway hyper-responsiveness) potentiated by frequent lower respiratory tract bacterial infections in childhood [Murphy TF, 2000; Sethi S, 2000]. Chronic mucus hyper-secretion and poor control of lung elastase activity causing increased lung tissue damage both occur, which contribute to the development of chronic bronchitis and/or emphysema. All of these pathologic changes caused by recurrent childhood infections make the individual even more susceptible to the further lung damage caused by cigarette smoke and other pollutants.

## **2.1.3 Epidemiology**

### **2.1.3.1 Prevalence**

A recent study reported on the prevalence of COPD in Canada by patient age [Lacasse Y, 1999]. Notably, the prevalence of COPD is clearly age-dependent; based on clinical presentation and spirometry data, 5% to 10% of patients who are at least 55 years of age have COPD. Data from the early 1980s show that the prevalence of this disease is increasing in Canada [Petty TL, 1997; National Institutes of Health, 1996].

### **2.1.3.2 Mortality**

COPD has consistently ranked as the fourth leading cause of mortality and morbidity in Canada and the United States; and is the only cause of death that has increased in prevalence over the last several years. For instance, the number of deaths attributed to COPD almost doubled between the early 1980s (4438) and the mid-1990s (8583) [Lacasse Y, 1999].

### **2.1.3.3 Morbidity**

Although many COPD exacerbations can be successfully treated in the outpatient setting, hospitalization is common. Individuals with COPD may need to be hospitalized for treatment of an acute exacerbation or in the final end stage of their disease. The total annual number of hospital separations—*hospital morbidity*—in which COPD was the primary diagnosis increased by 32% from 1981-82 to 1993-94 (from 42,102 to 55,785) [Lacasse Y, 1999]. Moreover, respiratory diseases, including

COPD, were the third most common primary diagnoses contributing to hospitalization in both genders in 1998 (13% of all hospitalizations for men and 11% of those for women) [Health Canada, 2001]. Hospitalization rates vary among age groups. For example, beginning at age 55 years, hospitalization rates for COPD in 1998 increased steadily with age among both men and women. Rates were higher for men than women particularly among the elderly [Health Canada, 2001]. It is also noteworthy that the rates of hospitalization are the highest in the eastern provinces—including Quebec [Health Canada, 2001]. The variation in hospitalization rates among provinces may reflect variations not only in prevalence, but also in the way health services are delivered.

#### **2.1.4 Burden of Chronic Obstructive Pulmonary Disease**

The real burden of COPD is best understood by comparing its indexes of mortality and (hospital) morbidity with other health conditions. Such data recently became available in a report detailing the causes of mortality and hospitalization among Canadian senior citizens. For the period from 1984 to 1993, COPD was the fourth-ranked cause of both mortality and morbidity in Canada [Stokes J and Lindsay J, 1996]. Therefore, it exerts a major economic impact on the Canadian health care system. The direct costs of health care associated with COPD, such as drugs, physician care, and hospital care, are estimated to be more than 3.5 billion dollars per

year [Health Canada, 2001; Health Canada, 1998]. The economic burden of this disease is more prominent if one takes into consideration the less visible or indirect expenses associated with COPD. For example, if one includes indirect costs associated with disability, work absenteeism, and early retirement, the economic burden would be even more significant (more than 12 billion dollars) [Health Canada, 2001; Health Canada, 1998]. The proportion of direct health care costs attributed to respiratory diseases ranked in the top 5 in terms of total costs in both 1993 and 1998 reports [Health Canada, 1998].

These 3 factors: high smoking prevalence (by far the most common etiological cause of COPD), increasing COPD prevalence, and rising costs of this disease; converged on health care system can easily motivate Outcomes Research studies.

- **EFFECTIVE MANAGEMENT OF COPD:** Tracking trends in COPD outcomes over time needs to be carried out so that health care system can plan for future increases in the need for services in view of predicted increases in COPD. This need is especially urgent in countries with limited health care resources.
- **DATA COLLECTION:** Data are needed on the use, costs, and relative distribution of medical and non-medical resources for COPD, especially in countries such as Canada where smoking and other risk factors are prevalent. These data are likely to have some impact on health policy and resource allocation

decisions. As options for treating COPD grow, more research will be needed to help guide health care personnel and health budget managers regarding the most efficient and effective ways of managing this disease.

- **OUTCOMES ASSESSMENT THROUGH A ROBUST AND INEXPENSIVE METHOD:** While conducting research is recommended to assess and monitor COPD outcomes, some (statistical) techniques need to be developed and evaluated in epidemiologic studies. Reproducible and inexpensive methodologies (e.g., use of RAMQ database) need to be evaluated and their use encouraged. This can ensure economical and accurate measurement when a relatively inexpensive and time-saving administrative database is being used.
- **IDENTIFICATION THROUGH EFFECTIVE TECHNIQUES:** Since COPD is not fully reversible (with current therapies) and slowly progressive, it will become even more important to identify all cases of resource utilization, particularly hospitalizations, as more effective techniques in data analysis emerge. Data to reveal whether or not screening and performance are efficient in directing management decisions in COPD outcomes are required.
- **HEALTH CARE UTILIZATION:** The specific components of effective health care utilization for COPD patients need to be determined. It is not known, for example, whether management plans are effective in reducing health care costs (e.g., by reducing the LOS) or improving the outcomes of exacerbations.



Developing and evaluating effective tools for management of COPD will be important in view of the increasing public health problem presented by COPD.

## 2.2 Clinical Indicators in Outcomes Research

### 2.2.1 Introduction

Outcomes Research is rapidly becoming recognized as an essential component of health care systems. As providers continue to explore new technologies and develop more efficient ways of providing services to patients, Outcomes Research becomes even more of a necessity; and the search—to *find methods for measuring and monitoring Outcomes indicators in clinical settings and their impacts on quality of health care*—continues. It is driven by the recognition that health care resources are finite; but our aging population, with an increasing prevalence of chronic diseases, requires more health care. This means each part of the health care system must undertake responsibility for its own activities and should be able to justify its actions when another party is affected (i.e., accountability) [Emanuel EJ and Emanuel LL, 1996].

Hence, identification and assessment of a set of clinical indicators, which may be used to monitor and evaluate the outcomes and quality of acute clinical care provided within public hospitals, are greatly encouraged. In fact, Outcomes Research provides the ability to identify areas of risk and to assist decision-makers in directing resources and strategies. However, these tasks have many challenges and obstacles to be overcome. These obstacles can be summarized as:

**1. Ambiguity:** How to identify and measure the end results in health institutions by the use of available resources (e.g., RAMQ databases)? Generally, making use of new data *to assess the distribution of health care resources and the effectiveness of clinical practices* are highly required. Yet, measuring clinical indicators using a new data set is not always an easy task.

**2. The validity of data gathering and data infrastructure:** Therefore, the contribution of chance, bias, or confounding to a difference in outcomes cannot be excluded [O'Connor GT, 1993]. Moreover, it could be more difficult if the information regarding these indicators is not directly collected and must be explicitly deduced.

**3. Costs:** Although a few existing databases allow the measurement and evaluation of Outcomes indicators in hospital settings, the complicated process to access and legal hurdles for the verification can be costly in terms of time delays. This, in turn, can effectively prevent a large number of researchers from being able to readily conduct their studies. Consequently, assessment of Outcomes indicators through robust and inexpensive methods is highly recommended; implementation of new methods may lead to cost savings.

## 2.2.2 Background

Obviously, there is a need for objective tools, and a desire for development of algorithms, which will allow precise measurements and rational comparisons of hospital outcomes in health care systems. This would satisfy a multitude of users ranging from the providers of health care to administrators, consumers and governments [The Quality Indicator Study Group, 1995; Sisk JE, 1990].

Amongst the many proposed objective tools are clinical indicators, which are viewed as a guide for the evaluation of the management and outcomes of patient care. *“The purpose of a set of clinical indicators should be to provide objective, quantitative measures of the application in clinical practice and service delivery of evidence on the clinical and cost effectiveness of healthcare interventions”*[Gold L, 1998]. These measures may increase the awareness of the public regarding the policies of the government and health care facilities.

### 2.2.2.1 Definition: Clinical Indicator

There are a number of definitions for clinical indicator. It is often defined as *‘a measurable element in the process or outcome of care whose value suggests one or more dimensions of quality-of-care and is theoretically amenable to change by the*

*provider*' used in clinical management and decision-making [Bernstein SJ and Hilborne LH, 1993]. Simply stated a clinical indicator provides a means to reflect on and improve clinical practice, it is not a precise measure of quality of care; rather it is a screen or flag which indicates areas for more detailed analysis.

Clinical indicators can be divided into groups based on:

- (1) whether the clinical indicator measures a process or outcome of care;
- (2) which domain of quality the clinical indicator is designed to measure; and finally
- (3) whether the clinical indicator can be constructed into a rate-based measure or if it should be regarded as a sentinel event.

#### **2.2.2.2 Process or outcome indicators**

*'The large numbers required to identify statistical differences in outcome measures and the confounding effect on outcome of factors such as severity, casemix and demography are likely to make process indicators more attractive'*  
[Gold L, 1998].

It is an important debate whether a clinical indicator is the outcome of care or the process of care; and in which context it is being measured. For instance, whether it is an evaluation of the technical performance of medical care for COPD by comparing the total number of patients treated, (1) with the number of patients who were hospitalized (**outcome**), or (2) with the number of patients who received the appropriate therapy (**process**).

Unless the crucial link between the process and outcome of care is established this debate cannot be resolved. At the center of the debate is the capacity of the human body to accommodate poor care and still have a good outcome and the fact that excellent medical care cannot guarantee against a poor patient outcome.

#### **2.2.2.2.1 Outcomes of care**

Outcome measures are often seen as the 'Holy Grail' for quality measurement [Berwick DM, 1989]. Generic markers of patient outcomes are commonly used as surrogate measures of quality-of-care because of the relative ease of obtaining this information, including **mortality** and **morbidity indexes**. Morbidity is often described by the length of hospital stay, ratio of post operative wound infection, or ratio of hospital readmissions.

These outcome measures are perceived as more objective and clearly defined events with obvious relevance to the patient. The idea that the end result of care to the patient is the best indication of care is a simplistic and attractive concept [Donabedian A, 1989], while it is a fallacious one in practice [Caplan RA, 1991; Donabedian A, 1988]. Moreover, markers of an adverse event, a direct link of these outcomes to the quality of the hospital care delivered, cannot be easily made because of confounding variables such as patient and disease factors [Brook RH, 1996]. However, generic

outcome measures have been shown to be relatively efficient at addressing improvements in quality-of-care [Domingo-Salvany A, 2002; Fan VS, 2002].

#### **2.2.2.2.2 Process of care**

The potential advantage of process over outcome measures include the ability to readily identify the area of care deficiencies [Brook RH, 1996; Mant J and Hicks N, 1995; Brook RH and Appel FA, 1973], thereby, giving the efforts to improve care for these patients. An additional factor is that process of care (especially when judged using explicit criteria) could be more sensitive to poor care than outcome measures [Mant J and Hicks N, 1995]. This makes a strong case for using process rather than outcome measures when evaluating care particularly for medical as opposed to surgical care. The critical element, in this case, is to identify the processes of care, which are clearly proven to influence a patient's outcome, such as the use of mechanical ventilation in the treatment of COPD [Soto FJ and Varkey B, 2003]. If objective evidence does not support a particular process of care, selecting the process to measure becomes difficult. Further the relevance of the particular process of care to quality is debatable. As with outcomes, the interpretation of process measures requires information about the patient characteristics and disease stage (complexity).

### **2.2.2.3 Domains of quality**

Both process and outcome measures can be used to address different aspects of health care, from the purely technical aspects of clinical procedures in a hospital to the outcomes of preventive health measures. These domains of health care can be expressed in many ways. But they are essentially offering fair access to health services; providing whether health care is effective, appropriate, timely, and safe; conveniently demonstrating that the resources are used to achieve value for money; and ensuring that the health care system is sensitive to individual patients' needs.

### **2.2.2.4 Rate-based or sentinel event indicator**

A sentinel event is an 'unexpected, usually avoidable event that must be investigated because the consequence is so grave', such as post-operative mortality. These events are typically rare and so provide insufficient opportunities for improvement. However, sentinel events are a very important source of information if a detailed critical analysis is undertaken [Langford AM and Reitz JA, 1995].

Rate based indicators are expressed as a ratio with the number of occurrences divided by the total patients at risk for the event. These indicators can be used to data trend using either external comparative data or for internal monitoring [Langford AM and Reitz JA, 1995].



### **2.2.3 Common Themes**

The common themes of performance measures can be summarized as (1) specific evidence-based process measures and (2) generic or specific outcome measures. These estimates can be applied to a single episode of care (i.e. hospital stay) or more broadly to the population. But the majority of outcome-based clinical indicators often reflect short term (e.g., the period of a hospitalization) and usually a proxy outcome (e.g., readmission).

The common generic outcome measures are usually based on mortality and hospital morbidity such as length of stay and readmission. Specific outcome measures are determined by the different interventions available for different diseases that can be extremely diverse, which are beyond the scope of this thesis.

#### **2.2.3.1 Evidence-based process measures**

Clinical indicators developed from evidence-based medicine have an inherent attraction because these indicators usually readily satisfy the criteria for validity, clinical relevance and potential to improve. However, the greatest concern with these clinical indicators is the ability to extract regularly and efficiently the necessary data. These indicators are normally developed from clinical research trials and practice guidelines.

The validity of the majority of process-of-care clinical indicators is based on face validity and consensus opinion rather than randomized controlled clinical trials [Mannion R and Davies HT, 2002]. The major limitation of using these process-of-care clinical indicators is the need for an intensive and costly effort to collect the data from the medical records.

### **2.2.3.2 Generic outcome measures**

#### **2.2.3.2.1 Mortality**

Mortality is a critical event that requires continuous monitoring in all health care settings. Mortality data raises many questions and has become the focus of consumers, insurers and public health agencies [Iezzoni LI, 1997b; Garnick DW, 1995; Knaus WA, 1993; Green J, 1991; Seagroatt V, 1991; Kahn KL, 1988; Dubois RW, 1987b].

The clinical indicators about mortality can be considered in the following broad categories:

- 1) overall mortality rates (e.g. total in-hospital mortality)
- 2) mortality rates for specific conditions or procedures:

- i) high risk conditions (in-hospital mortality for patients with acute myocardial infarction [Garnick DW, 1995])
  - ii) elective low risk conditions (post-surgical mortality [Roos LL, 1990])
- 3) death within a specified time period
- i) period of greatest risk (mortality within 24 hours of surgical processes such as percutaneous transluminal coronary angioplasty [Mills TJ, 1989])
  - ii) during the entire in-hospital stay (in-hospital mortality following knee replacement [Mangaleshkar SR, 2001])
  - iii) 30 day survival and beyond (death within 30, 90 & 182 days of hospitalization for hip fracture [Roos LL, 1996b])

The use of mortality rates, however, is almost always linked with the issue of risk adjustment. The complex series of events leading to a patient's death are difficult to unravel; in apportioning blame (i.e., poor quality of care) the relative contribution of the patient/disease factors must be considered. Studies [Dubois RW and Brook RH, 1988; Dubois RW, 1987a] in the late 1980s examined the concept of screening hospitals for poor quality of care based on their mortality rates. Comparison between the expected and observed death rates were made, and if differences existed, the

question of whether the hospital performance or the uniquely ill patient populations was responsible for the deaths still had to be answered.

In 1986, the USA Health Care Financing Administration (HCFA) released risk adjusted hospital mortality rates for Medicare patients in an effort towards public accountability. Similarly the NHS also published «league tables» for several years of overall mortality rates. These efforts had a limited impact on improving quality of care and the HCFA eventually abandoned its program because the method for determining quality of care was not valid. These included concerns about coding inconsistencies, inability to control adequately for the severity of illness [Fleming ST, 1995; Hartz AJ, 1995]. Further, the reactions of hospital leaders to the HCFA data was extremely negative, the hospital leaders expressed major concerns about the accuracy, usefulness, and interpretability of the mortality data [Berwick DM and Wald DL, 1990].

Mortality rates for procedures (e.g., following cardiac surgery) have the most prominence in the debate about the use of these performance measures for quality. The time between a surgical intervention and death is usually closely linked; but this is not the case with chronic medical conditions (e.g., heart failure).

*Clinical significance*

All clinical practices and interventions entail some risk; however, mortality following common elective procedures for uncomplicated cases should rarely occur. By comparing mortality rates across institutions or regions and by observing trends in mortality rates for elective procedures, it may be possible to target areas requiring more in-depth analysis or quality improvement efforts. Post-operative mortality, although not necessarily a sentinel event, represents a highly significant outcome that is potentially related to quality of care.

In fact, mortality is a critical event that requires continuous monitoring in all health care settings. Moreover, mortality data raises many questions and has become the focus of consumers, insurers and public health agencies [Stokes J and Lindsay J, 1996; Fleming ST, 1995; Green J, 1991].

Mortality rates within 24 hours of a procedure may provide a measure of the appropriateness of case selection and the appropriateness and utilization of scarce resources [Hannan EL, 1995; Leape LL, 1993]. If the physician has determined that a patient's prognosis is likely to result in death, medical intervention (procedures) may not be in the best interest of the patient or the best use of resources.

Furthermore, mortality during hospitalization can be expected in a certain percentage of cases, and this percentage is often higher when the patient population is

"sicker" [Librero J, 1999; Jencks SF, 1988b], having more complications and comorbid conditions than a comparable group.

There are ample citations in the literature that permit prediction of credible mortality rates that result from inexplicable and uncontrollable clinical problems unrelated to the surgical procedure itself. These published rates permit a facility to compare its similarly risk-adjusted rate in order to detect potential quality of care problems. The interpretation of a surgical mortality rate that has not been risk-adjusted may give an inaccurate and misleading impression of the hospital's standing in terms of quality of patient care.

#### **2.2.3.2.2 Inhospital Length of Stay (LOS)**

The current health care environment requires efficient use of hospital resources. Time spent in a hospital should be productive and cost-efficient for both the provider and the consumer, while ensuring that the quality of care is not jeopardized [Marcin JP, 2001].

Length of stay (LOS) can be used as a proxy of consumption of resources and costs in health care settings, particularly in the absence of cost data. Generally, LOS is a marker of cost-efficient utilization in both hospitals and intensive care units. It can be used as a surrogate for efficiency and in a principal target for cost containment in

modern health care analysis [Weingarten S, 1998; Pollack MM, 1994; Mozes B, 1991; Loop FD, 1983].

LOS is also influenced, however, by severity of illness, case complexity, and other case-mix variables such as patients' demographics (e.g., medical history, age, physical and mental health status, etc) [Ruttimann UE, 1998; Ruttimann UE and Pollack MM, 1996; Knaus WA, 1993; Epstein AM, 1990; Pollack MM, 1987; Berki SE, 1984; Goldfarb MG, 1983]. Goldfarb et al. (1983) are typical in finding that patient age and severity of illness are very important determinants of LOS. Moreover, Berki et al. (1984) acknowledged that LOS is positively associated with case complexity and severity, by which high variability in LOS can be explained up to 65%. In another study, Epstein et al. (1990) examined the impact of income, education, and occupation on LOS, and found that patients of lower socio-economic status have longer lengths of stay in the hospital.

It can also be considered as a surrogate measure of comorbidity in hospitalized patients. Previous investigations have shown that the behaviour of length of in-hospital stay changes depending on the levels of chronic comorbidity [Librero J, 1999; Romano PS, 1993b; Deyo RA, 1992], in which the length of stay, standardized by age and gender, significantly increased with each level of the comorbidity index. The proportion of readmissions was also associated with the gradual increase of the comorbidity score, in these studies. In general, greater comorbidity is associated with longer LOS and a higher proportion of readmissions.

LOS can also be considered as a performance indicator. Generally, refinement of performance indicators, such as LOS, is one of the areas of activity that can advance Outcomes research. These indicators need to be captured consistently and accurately if they are to be incorporated into outcome studies. Much time is spent on clearly defining each indicator and defining efficient processes for obtaining and reporting the data. Outcome reports are helpful in preparing clinicians and managers to look at data and consider a variety of parameters in assessing their services.

LOS can be feasibly obtained as the difference, in days, between date of discharge and date of admission in the index episode. Discharges on the same day can be considered as 1-day stays. It can also be calculated as an average (*ALOS*), that is the total number of days spent in the hospital by a given set of patients, divided by the number of discharges. The ALOS affects charges because longer stays generate higher charges. LOS may also be presented as *median*. The median represents the middle value of a distribution; half the values lie at or above the median, and half lie at or below it. Average (mean) measures can be significantly affected by a few unusually low or high values (“outliers”). Medians, on the other hand, are not affected to such a degree by outliers. The median, which is also called the 50<sup>th</sup> percentile, may be more representative of the typical length of stay than the mean.



### *Clinical significance*

LOS is generally related to the severity of the patient's *clinical* state and to external factors beyond the influence of a hospital. Prolonged lengths of stay may be an indication of clinical complications. In many cases, the length of stay can serve as a surrogate for the efficiency of care for a particular diagnosis. Extended lengths of stay can result from a variety of causes ranging from selection of "sicker" patients, to insufficient support services (e.g., radiology, laboratory), to substandard medical care or lack of resources for discharge (e.g., no funds for contract nursing home replacement, non-supportive family).

Moreover, comorbidity influences many different outcomes of hospital care such as LOS [Librero J, 1999; Greenfield S, 1993; Deyo RA, 1992; Cleary PD, 1991], and quality of life [Katz JN, 1994; Lahad A and Yodfat Y, 1993; Deyo RA, 1992; Linn BS, 1968]. Preoperative length of stay may reflect inefficient scheduling, which has a potential impact on the quality of patient care and hospital management. For the elective procedures, the preoperative work-up may be scheduled well in advance of the admission. The shortest inpatient stay possible presents the least opportunity for exposure to nosocomial infections and other untoward events.

The length of stay indicator has been excluded on the basis that it does not reflect the quality of care dimension in terms of clinical effectiveness. These indicators are used as proxy measures of cost and hospital efficiency. The length of

stay can be used as a screening test to identify outliers and perform a critical event analysis using medical record.

Finally, one of the principal benchmarks of inpatient efficiency has been the LOS required for a particular episode of care. Direct and indirect interventions aimed at reducing the LOS have included the development of different pathways of health care, as well as the institution of health care policies [Angus DC, 1996], thereby, lowering costs and improving care. In fact, a reduction in the LOS may indicate that clinicians are making use of procedures, such as day case surgery, which shift the burden of care onto community services and family after the patient spell.

#### **2.2.3.2.3 Readmission to hospital**

Clinical curiosity and the high economic cost of patient readmissions have stimulated the interest in hospital readmissions over the past decades. Research studies were initially facilitated by the ease and relative small costs associated with extracting the data from existing information technology systems such as administrative databases. These factors and the apparent face validity of hospital readmissions as a quality-of-care measure promulgated the use of readmission rates as a clinical indicator [Ibrahim JE, 1999; Ashton CM, 1997; Anderson GF, 1989].

The performance of hospital readmission as a marker of the quality-of-care has also improved, as the definition was refined from 'all readmissions to the same

facility' to 'an unplanned readmission within a certain period of separation (most often 30 days) to any hospital'. The latter definition was better able to exclude most of the readmissions unrelated to quality-of-care factors.

A comprehensive appraisal of the published research uncovered a small number of well-designed clinical based studies. The evidence to support the validity of readmission as a marker of quality-of-care was sparse due to methodological problems inherent in this field. Specifically, the use of multiple definitions of what constituted a hospital readmission and how the quality-of-care was defined and measured [Ibrahim JE, 1999; Milne R and Clarke A, 1990]. The largest and most convincing clinical trial [Ashton CM, 1995] indicated a small but significant proportion of unplanned readmissions were related to remediable quality-of-care factors in the index hospitalization. A subsequent meta-analysis also supported this finding [Ashton CM, 1997].

Direct translation of the validity of an unplanned readmission, as an indicator of quality to a rate based measure is not a simple step. A rate introduces other facets which must be considered, specifically the sensitivity and specificity of the variable (e.g., what proportion of patients are readmitted to another hospital) and adjustment of raw rates to account for confounding factors. These factors need to be understood and measured if accurate and appropriate comparisons are to be made between hospital readmission rates. They include post discharge mortality within time frame of

readmission, unplanned readmission to another hospital, the quality of the primary care, bed availability, and finally the appropriateness of hospital admission.

- 1) **Post discharge deaths within time frame of readmission:** The majority of clinically based readmission studies could not collect the information required to exclude patients who died prior to readmission. The death of these patients may be directly related to the previous hospital care [Ashton CM, 1997].
- 2) **Unplanned readmission to another hospital:** The number of patients who might seek treatment at different hospitals, especially if there have been quality-of-care problems, is likely to be significant. In Canada, 7.3% of patients who had heart attacks were readmitted to another hospital [Canadian Institute for Health Information, 2002], while in the United States a much higher rate, 13.8%, has been observed in a retrospective chart review [Beggs VL, 1996].
- 3) **The quality of the primary care:** The quality of primary (i.e., outpatient based) medical care [Bigby J, 1987] suggested 8.6% of emergency admissions were judged to be preventable as they related to lack of patient compliance or iatrogenic causes. There is increasing evidence to suggest that additional home based care can delay and prevent readmission [Stewart S, 1998; Rich MW, 1995] but this is not a consistent finding [Weinberger M, 1996].

- 4) **Bed availability:** Fisher and colleagues [Fisher ES, 1994] used the Medicare claims data to study the relative rate of readmission within 6 months in two demographically similar communities. They compared Boston with New Haven for five clinical conditions. A higher rate of readmission was observed in Boston (relative risk of 1.64, 95%CI: 1.53-1.76) which has a greater bed occupancy rate of 4.5 beds/1000 persons as compared to New Haven with 2.9 beds /1000 persons.
- 5) **Appropriateness of hospital admission:** A number of studies in Europe and the United States of America have examined whether hospital admissions are appropriate [Attena F, 2001; Esmail A, 2000; Pahor M, 1996; Bare ML, 1995; Lang T, 1995; Tsang P and Severs MP, 1995; Ludke RL, 1990]. The studies observed rates of inappropriate admission to be 9.1-31%. A large number of the admissions/readmissions were considered inappropriate because the care could have been provided on an outpatient basis.

### *Clinical significance*

As previously stated, there is a great likelihood of an association between the quality of care during a prior stay, adverse events, and treatment related complications. A prototype is an association between the LOS of an admission with an increased incidence of rehospitalization within a certain period of time. It also raises issues of adequate discharge planning, premature discharge, availability of adequate resources, and/or levels of patient compliance with discharge plan.

Moreover, unplanned readmission has face validity as a clinical indicator and is a significant issue consuming significant health care resources. The validity of this indicator as a quality-of-care measure remains questionable. It is clear that deficits in care can explain a small proportion of unplanned readmissions. Specific consideration, however, should be given to determining whether the readmission is unexpected or unplanned.

Comorbidity can also influence hospital readmission [Roos LL, 1991; Anderson GF and Steinberg EP, 1985; Anderson GF and Steinberg EP, 1984], although results are contradictory with regard to this last parameter [Waite K, 1994]. It is yet a surprising lacuna that only a few studies have shown an association between health utilization and readmission, which is a foreseeable proxy of comorbidity [Librero J, 1999; Waite K, 1994]. In other investigations, the previous use of hospital services has nonetheless been consistently associated with readmission [Corrigan JM and Martin JB, 1992; Evans RL, 1988; Fethke CC, 1986; Smith DM, 1985]. However, given the importance of readmission in the use of health services, researchers should consider this valuable parameter in utilization studies.

## 2.3 Data Collection and Sampling

Basically, a set of clinical indicators is required to address the current needs of the health care system. This would allow the opportunity for a more comprehensive coverage of clinical specialties and dimensions of quality. It may also limit the ability for gaming and the development of perverse incentives. However, this requires greater efforts to collect, analyze, and interpret the clinical indicators potentially diffusing and confusing organization's efforts to improve care [Gold L, 1998].

Currently there are three main methods for identifying clinical indicators—by direct examination of the *clinical medical record*, through an *administrative database*, or a combination of both or other *local in-hospital databases*. Other sources of data such as patient surveys (e.g., quality of life surveys, and patient satisfaction) are beyond the scope of this dissertation.

Identification of clinical indicators using clinical databases is considerably more labor intensive and requires health care personnel to review each hospitalization using the medical record, and then determine how to apply the clinical indicator definition. Although it is a complex system to maintain and is usually restricted to formal research studies, it allows the clinical indicator definition to be more detailed. Thus it is potentially better at identifying the patients with a deficit in their

quality-of-care. A typical example is a specific disease register for acute myocardial infarction [McCord J, 1992].

However, administrative databases can also be used to assess clinical outcome indicators. They are easily accessible and provide the opportunity for analyzing large sample sizes with the possibility of longitudinal follow-ups. Therefore, they could be used more efficiently in research projects. This source of data is more extensively explained in the following section.

The final approach is a combination of the two data sources. The administrative database is first used as a filter, and the selected patient files are subsequently reviewed to ensure compliance with the requested clinical indicator definition.

### **2.3.1 Administrative Databases**

Over the last few decades, the use of computers has significantly contributed to improve the process of data collection and storage. In fact, it allowed to collect a large volume of information concerning health services to be centralized and analyzed [Potvin L and Champagne F, 1986]. This granted the activity of individual health care institutions [Dubois RW, 1987b] and whole geographic areas to be characterized and quantified [Wennberg J and Gittelsohn, 1973].



The administrative databases contained a collection of patient demographics, summary codes of clinical events (diagnosis and procedure codes) and information for institutions and billing requirements [Wray NP, 1995; Romano PS, 1994; Romano PS, 1993a; Fleming C, 1992; Roos LL and Brazauskas R, 1990]. A number of well recognized databases have been used for the study of outcomes in health care including: The Oxford Record Linkage Study (United Kingdom) [Seagroatt V, 1991; Henderson J, 1989a; Henderson J, 1989b]; the Manitoba and the Saskatchewan Databases (Canada) [Harrison ML, 1995; Romano PS, 1994; Roos NP, 1986]; Health Care Financing Administration database, Veterans Affairs, and Medicare databases (United States); [Fleming C, 1992] and the Victorian Inpatient Minimum Data [O'Hara DA and Carson NJ, 1997; MacIntyre CR, 1997b] and the New South Wales Inpatient Statistics Collection [South M, 1997] (Australia).

These databases are capable of generating information regarding patient outcomes such as **length of stay**, **readmission**, and **mortality**; they can provide the opportunity for comparative analyses through the use of summary statistics to identify trends over time and between institutions.

Administrative databases are increasingly more accessible; they are the most popular source of data. Usually an algorithm written by an information technologist is used to identify the clinical indicator definition, which in turn is applied to all the separations within a defined time period. A separation episode is referred to an administrative process, by which a hospital records the completion of treatment and/or

care and accommodation of a patient: discharge, transfer, or death [Roos LL, 1996b]. The level of sophistication usually depends on the information contained in the database, the programming capacity, and the programmer level of competency.

Although administrative databases are a very valuable source of information for research on the outcomes of hospital care [Roos NP, 1996; Roos NP, 1995; Roos NP, 1988; Connell FA, 1987], any study conducted on the basis of information collected from administrative databases presents specific drawbacks. Most pointedly, this is because the cases under examination have not previously been assigned to random comparative groups. This aspect must be taken into consideration when information of this nature is used to compare the effectiveness of different technologies or the outcomes achieved by different providers. The confounding effect implied by different levels of severity of the patients' conditions also creates certain drawbacks when using administrative database [Librero J, 1999; Iezzoni LI, 1992; Iezzoni LI, 1990a; Greenfield S, 1989]. Another pitfall can be the presence of clinical evidence that an event had actually happened. For example, whether confirmatory evidence was present in a case of acute myocardial infarction (e.g., symptoms, laboratory, and diagnostic test data), rather than confirmation of timing and events from the perspective of ICD-9 codes in the database).

When attempting to use information from administrative databases for purposes of comparison, the researcher in fact encounters considerable methodological and conceptual problems. For instance, "severity" is a complex

construct that includes a wide range of clinical dimensions such as the severity of the main diagnosis, number and severity of coexistent diagnoses and complications, antecedents, age, physiopathological stability, and functional status. A number of non-clinical aspects associated with the prognosis must be taken into account, as well. These include the patient's socio-economic status, life style, and other relevant variables [Iezzoni LI, 1994].

Complicating this situation are the questionable quality of diagnostic data [Green J and Wintfeld N, 1993; Hsia DC, 1992; Roos LLJr, 1982], the shortcomings of diagnosis coding systems [Iezzoni LI, 1990b; Feinstein AR, 1988; McMahon LF, Jr. and Smits HL, 1986] and difficulties in establishing biostatistical models or algorithms on the basis of information found in these databases. Often, utilization of administrative databases has been criticized widely for a lack of precision and it is generally accepted that an error in the vicinity of 10-20% will be present in any health care administrative database [Iezzoni LI, 1994; Iezzoni LI, 1990b]. All of these problems have led to generating much controversy with regard to the usefulness and the limitations involved in using administrative databases for the purpose of making evaluations and comparisons [Hayward RA, 1993; Kassirer JP, 1993; Sanazaro PJ and Mills DH, 1991; Green J, 1991; Kahn KL, 1988; Jencks SF, 1988a].

However, tightened government spending and increased demands for health services have created a need to develop cost effective methods for evaluating medical care. For example in Canada, as the provinces struggle to meet service utilization

demands, researchers are examining the information contained in administrative databases. In some Canadian provinces, databases contain virtually all records of hospitalizations and physician visits for an entire population, enabling examination of rare events and exposures. With the data being already collected, research is less time-consuming than equivalent work involving primary data collection.

The remainder of this chapter describes the structure and the use of administrative databases in Canada and reviews what is known about the availability and accuracy of these claims database. Moreover, examples of the studies in which the quality of information in Canadian databases has been assessed will be provided.

### 2.3.2 Canadian Administrative databases

In Canada, administrative databases were initiated by provincial governments shortly after the advent of universal medical insurance to track hospital discharge summaries, physician billing claims, medical and pharmaceutical (prescription drugs) claims, and other health related data [Williams JI and Young W, 1996b]. Fueled by the revolution in computer technology, particularly concerning data management and advancements in information technology, these data have been used by researchers to study the epidemiology of diseases, physician practice patterns and health outcomes research. The scope of research embodied in the latter is wide and includes research on health outcomes and utilization variations across regions, access to care and cost efficiency.

Information potentially available in these databases can be classified into three domains. The first is *demographic* information, which includes but is not limited to patient age, sex, and place of residence.

The second domain is disease and includes *primary diagnosis*, *comorbidities*, and *treatment complications*. The definitions of diseases in most countries are based on the International Classifications of Disease (ICD); the 9<sup>th</sup> edition of the ICD is the current standard, known as ICD-9 [Practice Management Information Corporation, 2001].

The third domain relates to diagnostic, medical and surgical *procedures*. Unlike the ICD-9 codes adopted by all nations when reporting national health statistics, the procedure codes used in Canada are not uniformly coded. Some are based on the United States' Clinical Modification of the ICD-9-CM while others use the Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures (CCP), developed by Statistics Canada. In general, the varied use is merely a minor problem as the ICD-9-CM and the CCP are similar in construct; translation between the two coding systems is achieved with conversion tables available from the Canadian Institute of Health Information (CIHI) [Williams JI and Young W, 1996b].

Perhaps due to the above-mentioned benefits and the growing number of studies supporting the validity and reliability of the research, the administrative database research methodology is being applied more widely. However, a critical aspect of the validity of the research is the accuracy of the data within the databases.

### **2.3.3 Accuracy and Sources**

Data quality reflects both completeness of a dataset and its accuracy; the quality of administrative data presents ongoing concerns. Several American reviews of the quality of administrative data have criticized the use of these databases [Hannan EL, 1997; Ray WA, 1997; Iezzoni LI, 1997a; Romano PS and Mark DH, 1994; Fisher ES, 1992; Demlo LK, 1978]. Canadian databases, however, differ substantially from

those in the United States. Historically, financial incentives for recording inaccurate diagnoses have been minimal. However, in two Canadian provinces (Ontario and Alberta) the introduction of case mix adjustments for hospital global budgets has created incentives similar to those in the United States. Canadian efforts at data standardization now permit separating comorbidities from inhospital complications, which are not available in most American datasets. Accordingly, this can help generating histories of hospital and physician use before an index hospitalization (to help deal with the underestimation of comorbidities in hospital discharge abstracts), and examining mortality and hospital readmissions anywhere in a province (compared with just inhospital mortality or readmission to the same hospital in more limited datasets) [Jollis JG and Romano PS, 1998; Roos LL, 1997; Romano PS and Mark DH, 1994; Canadian Institute for Health Information (CIHI), 1994].

Administrative databases vary widely in the amount of research on their quality. Little has been published on registries, despite their central importance in building integrated data systems. Great attention has been given to hospital discharge abstracts, less attention to physician claims, and almost no attention to nursing home data. Studies assessing pharmaceutical data are just beginning [Melnychuk D, 1993]. This part comprehensively reviews studies examining the quality of various administrative databases, including hospital discharge abstracts, physician claims, and pharmaceutical data. We collected all published papers on the quality of administrative databases generated as part of the provincial health care systems.

PubMed and Medline were searched using the keywords: administrative data, data quality, accuracy, reliability, and Canadian administrative data. A few studies were eliminated for mentioning data quality only peripherally.



## 2.3.4 Substantive Databases

### 2.3.4.1 Populations and Events

Canada's universal health care coverage should provide information for essentially an entire population when data collection is mandatory. A record of demographic and clinical information is produced for every patient discharged from a hospital. Hospital databases are highly accurate in recording information regarding individuals discharged from inpatient stays [Kozyrskyj AL and Mustard CA, 1998; Tamblyn R, 1995; Mustard CA, 1995; Davidson W, 1994; Scherer K, 1994; Roos LL, 1993]. For example, Mustard et al. (1995) found maternal delivery records to account for more than 99% of Manitoba births from 1989 to 1991 (Table 2.1).

**Table 2.1:** Assessment of Population Coverage.

Primary Author, Date	Database	Years Examined	Population	Completeness of Coverage (%)
Davidson, 1994	New Brunswick Prescription Drug Program	1990-1991	100% residents	100%
Kozyrskyj, 1998	Manitoba Drug Programs Int'l Network	1995	100% residents aged $\geq 65$ , Not nursing homes	80% (aboriginal population); 93% (non-aboriginal)
Mustard, 1995	Manitoba Health Maternal Delivery Records	1989-1991	100% residents	99%
Roos, 1993	Manitoba Registry File	1989	100% residents	100%
Scherer, 1994	Manitoba Health Administrative Data	1986	100% residents	100%
Tamblyn, 1995	Régie de l'assurance maladie du Québec	1990	100% residents aged $\geq 65$	99%
Thiessen, 1990	Saskatchewan Drug Plan	1984	95% residents	100%

Fee-for-service payment generates a strong incentive for physicians to claim every patient visit. Movement away from fee for service to either lump sum payments to clinics or capitation-based charges may create data problems. Only one Canadian study has examined the accuracy of “evaluation claims” generated in such situations [Cohen MM, 1993]. Although, in this study, testing (Papanicolaou Smear) was appropriately recorded at one major Winnipeg clinic where physicians were on a global budget, questions have been raised as to the recording of influenza immunization by clinics. While entry of “evaluation claims” is part of the clinics’ contractual responsibility, monitoring must be sufficient to ensure compliance.

#### **2.3.4.2 Hospital Discharge Data**

Demographic data, procedures, and diagnoses are the most commonly examined variables in studies of hospital data quality. Comparisons to another source, such as physician claims, survey data, or pharmaceutical records, are frequently made (Table 2.2). Hospital discharge abstracts are often evaluated by reabstracting the hospital chart information onto a second discharge abstract, and then comparing the two abstracts (Tables 2.3, 2.4). Several Saskatchewan and Manitoba studies conducted both types of checks on the same data [Edouard L and Rawson NS, 1996; Rawson NS and Malcolm E, 1995a; Rawson NS and Malcolm E, 1995b; Malenka DJ, 1994].

**Table 2.2:** Comparison of Hospital Discharge Summary with an Additional Source.

1 <sup>st</sup> Author, Date	Records Examined	Years	Variable Examined	Source of Comparison	Agreement
Cohen, 1996	9,780 women participating in a National Breast Screening Study	1979-1992	Mammogram hospital claims outside of the NBSS	Self-report	21% Claim 100% No Claim
Edouard, 1996	1,537 hospital discharge abstract with a hysterectomy procedure	1 Calendar year	Procedure Diagnosis	Physician claims	56% Diagnosis 98% Procedure
Iron, 1995	568,741 hospital discharge abstracts of women with mammogram	1988-1992	Procedure	Physician claims	88% Hysterectomy 96% Cholecystect.
Rawson, 1995a	1,923 hospital discharge abstracts for cholecystectomy	1986	Procedure	Physician claims	81% Hysterectomy 98% Cholecystect
Rawson, 1995b	9,537 hospital discharge abstracts for heart disease and COPD	1986	Diagnosis	Physician claims	85-94% Heart Dx 80-95% COPD
Rawson, 1995	1,474 hospital discharge abstracts for schizophrenia or depressive disorder	1986	Diagnosis	Physician claims (SMHS)	62-89% Schizophrenia 67-93% Depression
Roos, 1982	3,131 hysterectomy and cholecystectomy hospital discharge abstracts	1974	Procedure	Physician claims	89% Cholecystect. 91% Hysterectomy
Roos, 1989	61,310 hospital discharge abstracts for 11 procedures	1979-1983	Procedure	Physician claims	94% Vascular 99.7% Valve replace
Roos, 1991	112,000 hospital discharge abstracts for cholecystectomy or prostatectomy	1979-1984	Comorbidity	Clinical dataset	65% Prostatectomy 92% Cholecystect.
Roos, 1996	1,955 hospital discharge abstracts for angiography	1989-1990	Procedure	Physician claims	98%
Tennis, 1993a	1,333 hospital discharge abstracts for rheumatoid arthritis	1978-1980	Cancer incidence	Prescription, medical, & cancer files	98%

**Table 2.3:** Summary of Canadian Reabstraction Studies of Hospital Records.

Data Source, Date	Province	Abstracts	Years	Type of Data	Agreement
<b>General Studies</b>					
Hospital Association, 1991	ON	2,998 records (43 hospitals)	1988-1989	Demographic Diagnosis Procedure	93-100% 37-81% 53-95%
Newfoundland, 1995	NF	850 records (6 hospitals)	1994	Demographic Diagnosis Procedure	98-100% 57-74% 93%
The Doctor's Hospital, (Toronto) 1992	ON	300 records (1 hospital)	1988-1992	Demographic Diagnosis Procedure	100% 95% 96%.

With the exception of infrequently used variables such as postal code and family characteristics, demographic data on hospital discharge abstracts are very accurate (Tables 2.3, 2.4). Approximately, more than 95% agreement for demographic information has been found on the charts of both schizophrenia and depressive disorder patients [Rawson NS, 1997] and of knee replacement surgery patients [Hawker GA, 1997]. In a different study, slightly lower agreement levels were found: 84% for date of birth, 86% for marital status, and 100% for gender between charts and records of mental health service recipients [Robinson JR and Tataryn DJ, 1997]. The different criteria used by the two systems for recording postal codes appeared to lead to much lower correspondence (59%) for this variable. Education, employment, occupation, treaty status, and religion were optional demographic variables, and also had low agreement rates, ranging from 47-74%.

**Table 2.4: Summary of Canadian Reabstraction Studies.**

1 <sup>st</sup> Author, Date	Province	Abstracts	Years	Type of Data	Agreement
<b>Condition/ Procedure Specific Studies</b>					
Cohen, 1993	MB	61 records of Papanicolaou tests	1970-1984	Patient ID	95%
Delfino, 1993	QC	1,279 records (14 hospitals)	1984, 1987	Diagnosis	95% Asthma 76% Others
Edouard, 1996	SK	226 records of hysterectomy	One calendar year	Admin. Data Diagnosis	95% 85%
Hawker, 1997	ON	175 records of knee replacement (CIHI)	1984-1990	Demographic Procedure Complication	> 94% > 94% 66-100%
Humphries, 2000	BC	817 records of PCI (1 hospital)	1994-1995	Comorbidity	63-88%
Malenka, 1994	MB	485 records of prostatectomy	1974-1980	Comorbidity	93-98% $\kappa = 0.14$ - 0.86
Pinfold, 2000	ON	735 records of breast cancer w/wt mastectomy	1991	Procedure	85-98%
Rawson, 1995a	SK	444 records of cholecystectomy or hysterectomy	1986	Demographic Diagnosis Procedure	100% 88% 100%
Rawson, 1995b	SK	AMI, 224 records COPD, 225 records	1986	Diagnosis: AMI COPD	97% 94%
Rawson, 1997	SK	281 records of schizophrenia or Depression (DD)	1986	Demographic Schizophrenia DD	95% 77-94% 58-94%
Ray, 1989	SK	236 records of hip fracture	1984-1985	Diagnosis	99%
Robinson, 1997	MB	140 records from MMHMIS	1994-1995	Demographic Diagnosis Contact Data	60-100% 56-83% 96%
Tamblyn, 2000	QC	1,181 records for injuries (10 hospitals)	1993-1994	Date Expanded date Diagnosis	67% 81% 14-97%
Tennis, 1993b	SK	1,717 records, 150 with Rheum Arthr	1970-1987	Date Diagnosis	95% 83%
West, 2000	SK	Stratified random sample of 600 individuals	1994-1995	Diagnosis	75-77%

**Note:** BC: British Columbia; MB: Manitoba; ON: Ontario; SK: Saskatchewan; QC: Quebec

Accuracy of hospital discharge abstracts for procedures performed range from good to very good (Table 2.4). The agreement between hospital discharge abstracts and physician claims from 1974 were 89% and 91% for cholecystectomy and hysterectomy, respectively (Table 2.2) [Roos LL, 1982]. More recently, Iron et al. (1995) found agreement to range from 88% for hysterectomy to 94% for breast surgery and 96% for cholecystectomy, while Rawson and Malcolm (1995a) found agreement levels of 81% for hysterectomy and 98% for cholecystectomy. In Roos et al. (1996), hospital discharge abstracts and physician claims for coronary angiography between 1989 and 1990 were linked with a successful matching rate of 98%.

Both the diagnosis and its importance to the hospitalization influence accuracy of recording on hospital discharge abstracts. Ray et al. (1989) found 235 of 236 patients with hip fracture to have a diagnosis of hip fracture on their charts. Relatively good levels of agreement have also been reported for various respiratory diseases, with numbers ranging from 80-97% [Rawson NS and Malcolm E, 1995b]. Delfino et al. (1993) reabstracted hospital discharge records, with agreement of 95% for asthma, 76% for non-asthma respiratory diagnoses, and 93% for non-respiratory diagnoses (Table 2.4) [Delfino RJ, 1993].

Lower rates of agreement were recorded between hospital discharge abstracts and charts for some psychiatric disorders; Rawson et al. (1997) found 94% agreement for diagnosis of schizophrenia, but only 58% for depressive disorder. Agreement with physician claims was more consistent, at 89% for schizophrenia and 94% for

depressive disorder [Rawson NS, 1997]. Generally, diagnoses other than the primary or "*most responsible*" diagnosis generally show less agreement. Agreement among all mental health services recipients was 76% on primary ICD-9-CM diagnosis at the 5-digit level and 83% on 3-digit level diagnosis [Robinson JR and Tataryn DJ, 1997]. Agreement on comorbid diagnoses was 55% and 57% at the 5-digit level and the 3-digit level, respectively.

In a Manitoba study of comorbidities associated with prostatectomy, agreement rates ranged from 88%, with a chance-corrected agreement (Kappa,  $\kappa$ ) of 0.56, for chronic obstructive pulmonary disease to 99% ( $\kappa = 0.78$ ) for tumors [Malenka DJ, 1994]. Roos et al. (1991) focused on prostatectomy and cholecystectomy records from 1979-1983, with agreement levels similar to those in Malenka et al.'s study [Roos LL, 1991]. As noted earlier, the rarity of a comorbid condition tends to increase percentage agreement between two data sources. Use of the Kappa statistic instead of percentage agreement shows considerable room for improvement in abstracting from patient charts to hospital discharge abstracts.

Although information on hospital discharge abstracts matches with that from physician claims with relatively high accuracy, reabstraction studies generally show somewhat lower rates of agreement. The numbers are still quite high when demographic data or procedures are examined, but they vary more than the agreement levels of other kinds of variables. Improved abstracting of information on comorbidities and complications from hospital charts is desirable. More information

on comorbidities can be generated by looking at hospitalizations and physician claims in the year before the index event. Although useful for case identification and estimating prevalence, the additional diagnostic information does not appear to markedly help in predicting outcomes [Roos LL, 1996a; Romano PS, 1993b].

### **2.3.4.3 Pharmaceutical Data**

Pharmaceutical or dispensed prescription claims data are useful for describing a population's medication use, both geographically and over time. The impact of changes in drug benefit plans, health outcomes after changes in drug therapy, and adverse effects of specific drug therapies can all be assessed using pharmaceutical data.

A provincial prescription claims database typically covers a select portion of the population, most commonly those aged 65 and over. Only Saskatchewan and Manitoba have databases relevant for the entire population [Miller E, 1996]. Saskatchewan data, however, excludes the drug utilization information of First Nations residents, making it impossible to generalize findings to the province's entire population.

Although Manitoba's First Nations population is included in the provincial Drug Programs Information Network, their information is not as complete as that for



other provincial residents. Manitoba prescription data was relatively complete when program participation was mandatory; for this population, 93% of events recorded in a manual audit were noted in the computerized database. Agreement was slightly lower among those for whom participation was optional (90%). Information on First Nations was noted in only 80% of the audit [Kozyrskyj AL and Mustard CA, 1998].

No provincial pharmaceutical database records drug dispensation in hospitals, nursing homes, or outpatient clinics. Also missing from prescription claims databases are information on patient allergies [Miller E, 1996]. Studies of nonprescription drugs (over-the-counter medications) cannot be conducted. Although persistence to treatments (or therapies) can be estimated, patient compliance can hardly be assessed using pharmaceutical data; biological indicators remain a better measure of drug exposure.

The defined daily dose (DDD), a commonly used measure of dosage level, depends on a database indicating the number of days of medication dispensed, a variable not included in every province. It also depends on the strength of the drug per tablet, making it impossible to calculate the DDD for non-solid drugs such as insulin, inhaled asthma medications, and some pediatric medications [Metge C, 1999].

With no diagnostic information available in prescription claims databases [Lewis NJ, 1993], it is difficult or impossible to determine the condition for which many drugs were prescribed using pharmaceutical data alone [Kephart G, 1995].

However, when linked with data such as physician claims records; pharmaceutical data is useful in studying various drug outcomes.

The few studies conducted on the quality of pharmaceutical data indicate generally high reliability (Table 2.1) [Tamblyn R, 1995; Davidson W, 1994; Thiessen BQ, 1990]. Tamblyn et al. (1995) found Quebec's pharmaceutical services database to be at least 99% complete on all fields. Davidson et al. (1994) reported that the New Brunswick Prescription Drug Plan covers 100% of those aged 65 or older in the province. Thiessen et al. (1990) made a similar report for the Saskatchewan Prescription Drug Plan, excluding First Nations.

#### **2.3.4.4 Physician Claims**

As seen earlier, physician claims are often compared to hospital discharge data to assess the reliability of the inhospital data. Although little work has specifically assessed their quality, physician claims have also been compared with survey responses and clinical measures to assess the reliability of the physician claims themselves. The variable most often examined in these studies is diagnosis.

Four available studies focused on Manitoba physician claims; two reviewed a combination of hospital discharge abstracts and physician claims (Table 2.5). Young et al. (1991) found 95% agreement between diagnosis of diabetes on hospital

discharge abstracts or physician claims and self-reports of diabetes from the Manitoba Longitudinal Study on Aging [Young TK, 1991]. Robinson et al. (1997) compared hospital and physician files with the Manitoba Heart Health Survey, assessing diabetes, hypertension, high cholesterol, stroke, myocardial infarction, or other heart disease. Agreement ranged from 79% (0.40 Kappa) for elevated cholesterol to 96% (0.72 Kappa) for diabetes [Robinson JR, 1997]. Agreement for hypertension was 84% (0.59 Kappa); similar levels of agreement were found in comparisons between administrative and survey data [Muhajarine N, 1997].

**Table 2.5:** Studies on the Accuracy of Physician Claims in Canada.

1 <sup>st</sup> Author, Date	Records Examined	Years Examined	Diagnostic Variable	Agreement
Muhajarine, 1997	Physician claims for 2,275 survey respondents	1987-1990	Hypertension	82-85% ( $\kappa = 0.56-0.60$ )
Roberts, 1994	Physician claims for records from the MIMS	1980-1994	Service date Immunization	98% > 98%
Robinson, 1997	Hospital discharge abstracts and physician claims for 2,792 survey respondents	1974-1990	Diabetes AMI and its risk factors	96% ( $\kappa = 0.72$ ) 87% ( $\kappa = 0.38$ )
Young, 1991	1,227 hospital discharge abstracts and physician claims for diabetes	1971, 1983	Diabetes	95%

Properly assessing data quality with so few studies and such variability is difficult. Diabetes, for instance, tends to yield high levels of agreement, while

agreement between sources varies considerably for respiratory disease. Checks on physician claims should be a higher priority in data quality assessment.

#### 2.3.4.5 External Criteria

Both hospital discharge abstracts and physician claims records can be compared to external standards developed by nongovernmental organizations and to expert (clinician) judgment. These comparisons usually vary more and produce lower levels of agreement than those examining sources within the same province. For example, McLean et al. (1994) found 74% agreement between the sources of data for the diagnosis of Guillain-Barre syndrome in Ontario; however, the percent agreement was found to be slightly higher in Quebec (79%) [McLean M, 1994].

Particular data elements are often lacking on the abstracts, leading to missing information when comparisons are made. Tennis et al. (1993b) found that only 45% of patients diagnosed with rheumatoid arthritis (RA) on 1978-1980 hospital discharge abstracts met five or more of the American Rheumatism Association (ARA) criteria [Tennis P, 1993].

Agreement between administrative data and physician judgment is more important. When *diagnosis is relatively objective* and straightforward, as with diabetes, agreement levels are high [Young TK, 1991]. Agreement is lower with more

*subjective diagnoses* such as asthma [Sweet L, 1992] for which physicians may disagree. When diagnosis is uncertain, explicit criteria derived by physicians are crucial to improving diagnostic accuracy.

#### **2.3.4.6 Sensitivity Testing**

Sensitivity is important to data quality assessment; a test with low sensitivity will miss instances of the variable examined and produce inaccurately low levels of agreement. A very sensitive measure of a variable (a high proportion of instances identified) is likely to compromise specificity (a high proportion of non-instances excluded). The goal is a balance between high sensitivity/low specificity and low sensitivity/high specificity.

Several studies have found few important differences between two measures of a variable, a more sensitive algorithm versus a more specific one. The more sensitive algorithm to catch an event is usually an extension of the more specific one. For example, Romano et al. (1994) examining comorbidities among coronary artery bypass graft patients, found index hospital discharge abstracts (those from the hospitalization for surgery) and prior abstracts to be more sensitive than index abstracts alone [Romano PS, 1994].

### 2.3.4.7 Summary of Databases by Province

We examined work from all provinces in Canada; however, the main focus was on the province of Quebec which includes four studies [Tamblyn R, 2000; Tamblyn R, 1995; McLean M, 1994; Delfino RJ, 1993] assessing quality of data since 1981 (year of initiation of the Quebec administrative databases)<sup>1</sup>. A common method of assessing data quality in these papers is to examine the completeness of administrative databases by linking them with another source. Reabstraction and comparison between hospital discharge data and another source were also very common. Assessments on the quality of physician claims data are rare; with only four studies in Canada (all from Manitoba) noted. Two of these studies combined data from both physician claims and hospital discharge abstracts.

The quality of the provincial databases in Canada appears to be improving over time. New summary measures and case identification techniques should help improve our understanding of quality issues, but such assessments need to be performed more frequently.

---

<sup>1</sup> Pharmaceutical Services Claims: 1981; Hospital Discharge Data: 1983; Physician Claims: 1986.

## 2.4 Accuracy Assessment of Administrative Databases

The accuracy of administrative databases can be assessed using different techniques. The most efficient approach is to compare the information in one dataset to those collected in an independent dataset. This method, however, measures the *agreement* which is considered as a surrogate measure of *accuracy*, e.g., assessing the agreement of an episode of hospitalization or comparing a diagnosis in both datasets rather than the accuracy of either one. An example of this method would be to compare diagnostic codes from the hospital discharge abstracts (MED-ECHO database) to records archived in a physician claims database such as the one maintained by the Régie de l'assurance maladie du Québec (RAMQ) which is the main objective of this dissertation.

The difficulty lies in determining the gold standard dataset. If there is disagreement between the two datasets with respect to a defined variable, it is unclear in which database the inaccuracy lies. Similarly, if both databases being compared contained the same error, there would be agreement and this may falsely be determined to be accurate. However, if database information is collected independently, the frequency of this problem should be small.

The more widely accepted method is the chart audit but it is much more expensive and more time-consuming than the electronic data comparison. In this

process, data are extracted from the charts, usually by a medical record technician. The data are then entered into a database and compared with the electronic data records. A critical aspect of the validity of this process relies on the attending physician recording pertinent and accurate information [Lloyd SS and Rissing JP, 1985] and the abstractor being able to recognize this information. This latter point has been shown to be of little concern when trained medical record technicians are used to abstract the charts [Williams JI and Young W, 1996b].

The most rigorous but most time-consuming and restricted methodology is a comparison of the recorded data with expert clinical criteria. The method is restricted in that expert clinical criteria must be established for the specific diagnosis that is to be evaluated and it relies on these critical diagnostic elements being well recorded by the physician in the chart in order to assess the clinical criteria. However, this methodology allows investigators to examine not only the accuracy (between chart and database) but also the validity of the diagnosis. Below, a review of the agreement studies in Canada using different methodologies is shown.

### **2.4.1 Electronic Database Comparison Method**

There are eight Canadian studies identified which compare data in the hospital discharge summaries to those in physician claims databases [Rawson NS, 1997;



Edouard L and Rawson NS, 1996; Roos LL, 1996a; Iron K, 1995; Rawson NS and Malcolm E, 1995a; Rawson NS and Malcolm E, 1995b; Roos LL, 1989; Roos LLJr, 1982] or clinical databases [Chen E, 1995; Roos LL, 1991] maintained for other reasons (Table 2.2).

Demographic data were not reported in any of these studies. The agreement of the diagnosis contributing the most to the hospital stay was reported in one study where 61,310 claims for 11 different procedures were examined [Roos LL, 1989]. In this study, discharge summary data were compared to the Manitoba physician claims database. Linkage between hospital and physician claims data was excellent (over 95%) for 5 out of 11 surgical procedures (hysterectomy, prostatectomy, total hip replacement, coronary artery bypass surgery, and heart valve replacement); there was over 90% perfect agreement for three other procedures (cholecystectomy, cataract surgery and total knee replacement). Moreover, diagnostic agreement between physician claims and hospital discharge abstracts averaged 75% using three-digit ICD-9-CM codes, although it varied considerably across different diagnoses (from less than 20% to over 90%).

Secondary diagnoses are defined as comorbidities or complications incurred during admission. The agreement of secondary diagnoses listed on the discharge summary compared to data in a clinical database is best described in another study by Roos and colleagues [Roos LL, 1991]. These authors compared discharge summary data of 112,000 adult patients undergoing non-obstetrical procedures with data

collected in an anesthesia follow-up study. Overall, agreements ranged between 17-90% [Roos LL, 1991]. Most often morbidities were not recorded in the discharge summary.

## **2.4.2 Chart Audit Method**

### **2.4.2.1 Demographic Data**

There are fifteen Canadian reabstraction studies identified which have compared demographic, diagnostic, or procedural data from the hospital discharge records with information reabstracted from the chart by specially trained abstractors. Three studies from Ontario [The Doctor's Hospital, 1992; Ontario Hospital Association, 1991] and Newfoundland [Newfoundland Department of Health, 1995] examined random samples of all patients admitted to participating hospitals in each province; other studies examined patients admitted with specific diagnoses or procedures: cancer [Pinfold SP, 2000; Cohen MM, 1993], cholecystectomy and hysterectomy [Edouard L and Rawson NS, 1996; Rawson NS and Malcolm E, 1995a], depression and mental disorders [West SL, 2000; Robinson JR and Tataryn DJ, 1997; Rawson NS, 1997], hip fracture or other injuries [Tamblyn R, 2000; Ray WA, 1989], respiratory diseases [Rawson NS and Malcolm E, 1995b; Delfino RJ, 1993], prostatectomy [Malenka DJ, 1994], rheumatoid arthritis [Tennis P, 1993].

The largest reabstraction study reported is based on a random sample of 2998 inpatient records from 43 participating hospitals in Ontario from 1988-1989 [Ontario Hospital Association, 1991]. Overall 97% of the submitted admission and discharge dates matched the abstracted records, all deaths were correctly coded and 93% of the patients' postal codes matched. Another large study using a random sample of all patients admitted to 6 Newfoundland hospitals, also showed near agreement between electronic and abstracted data [Newfoundland Department of Health, 1995]. Of the 850 records abstracted, 100% of admission and discharge dates concurred while 98% of deaths were coded correctly. In this study, postal codes were not abstracted. The third study to reabstract random patient records, irrespective of diagnosis, was performed at a Toronto community hospital [The Doctor's Hospital, 1992]. Both inpatient and outpatient records were abstracted and there were no discrepancies in demographic information between the electronic and abstracted data.

Similarly, high accuracy rates for demographic information were found for studies comparing the chart data with the electronic records of patients with specific conditions or undergoing procedures. Agreement of various demographic data ranged between 84-100% for these studies.

Agreement of postal codes consistently has the poorest agreement within the demographic domain. Like the Ontario Hospital Association study, Delfino's study reported postal code agreement to be 84.3% compared to other demographic data which had agreements between 92-99% [Delfino RJ, 1993]. One may have predicted

this observation as this part of the demographic information would be expected to change during the course of a person lifetime (if they were to move their primary residence) and may not be updated in the electronic records.

Overall, the agreement between reabstracted and electronic sources of demographic data was very high, ranging from 84-100% and 92-100% if one excludes postal code data.

#### **2.4.2.2 Diagnostic Data**

In comparison, the agreement between reabstracted and administrative records is lower for diagnostic data than for demographic data. Several of the above studies examined the agreement between the reabstracted and electronic data records for the patients' diagnosis.

In the Ontario Hospital Association and the Newfoundland reabstraction studies, the abstracted *most responsible diagnosis*, matched the database code in 81% and 74% of cases, respectively. Reasons for the discrepancies include: incomplete documentation at the time of coding, hospital policy, and lack of specificity.

One set of investigators examined the problem of lack of specificity more closely. When Rawson and Malcolm [Rawson NS and Malcolm E, 1995a] reviewed

the charts of patients in large Saskatchewan hospitals who had a cholecystectomy, they found the agreement rate for the most responsible diagnosis to be 41.7% when the four-digit ICD code was used and 87.6% when the three-digit code was used. Similarly, patients who received a hysterectomy had agreement rates of 71.2% and 84.5%, respectively.

*Secondary Diagnoses*, in general, demonstrate even poorer agreement between abstracted records and administrative databases. For the Ontario Hospital Association reabstraction study [Ontario Hospital Association, 1991], secondary diagnoses showed agreement in only 37% of cases while it was slightly higher in the Newfoundland study at 56.8% [Newfoundland Department of Health, 1995]. While a poor overall agreement rate was reported for the random sample of charts reabstracted, not all hospitals in Ontario demonstrated uniformly poor secondary diagnosis agreement rates. Hawker found that of the comorbidities recorded on the chart, only 29.9% were recorded in the CIHI database while 66% of the complications listed on the chart were coded in the database [Hawker GA, 1997]. Furthermore, in Doctor's Hospital study [The Doctor's Hospital, 1992] which sampled 165 inpatient records over a 2-month study period, abstracted comorbid conditions and complications matched administrative data records 95% of the time.

### **2.4.2.3 Procedural Data**

Six of the above studies evaluated the agreement of the procedures recorded on the hospital discharge abstracts with data found in the clinical chart (Table 2.2). In the Ontario Hospital Association study, the agreement for the primary procedure was 88% while it was 53% for secondary procedures [Ontario Hospital Association, 1991]. The Doctor's Hospital study showed higher agreement in that center with a procedure agreement rate of 96% [The Doctor's Hospital, 1992]. Overall, the accuracy of the procedural data was found to be in an acceptable range.

### **2.4.3 Expert Clinical Criteria Method**

The expert clinical criteria method requires a sample of charts where the discharge summary was coded for a specific diagnosis. Moreover, there has to be expert clinical criteria derived for that diagnosis. The patients' charts are evaluated to see if there is an evidence to substantiate the diagnosis (according to the expert clinical criteria), recorded in each patient's chart.

The largest study of this type is the Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group [Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group, 1989]. A sample of patients from each province for the years

1977, 1981, and 1985 that were assigned the ICD-9 code 410 (acute myocardial infarction—AMI) on their discharge summaries were reviewed by trained clinical abstractors. The objective of this study was to assess whether patients with the code for AMI had been misclassified according to the World Health Organization (WHO) criteria. For the latter two years, patients with ischemic heart disease (ICD-9 codes 411, 414) were also assessed to determine whether patients had been misclassified according to the WHO criteria. Overall accuracy of the diagnosis was 80.5-89.2%, with increasing accuracy rates over time. In addition, in 1981 there was a 7.1% average reclassification rate for the two provinces of patients with ischemic heart disease that met the WHO criteria for AMI. Another study in Ontario [van Walraven C, 1990] also assessed the accuracy of AMI diagnosis on hospital separation records using the WHO criteria. The accuracy rate in their study was 79.4%.

In diseases where the diagnosis is relatively certain such as AMI, studies have confirmed high accuracy rates. However, for diseases that are less certain, accuracy rates are lower. An example of this observation is the study by Tennis and colleagues [Tennis P, 1993] who assessed the accuracy of hospital separation diagnoses in patients with rheumatic diseases based on the 5 rheumatoid arthritis (RA) criteria of the American Rheumatism Association. Of the 142 charts of patients diagnosed with RA, 45% of the charts had all five criteria, almost 21% had three to four, 24% met between one to two, and 7% had none of the criteria.

Thus far, different techniques to study the accuracy of health care administrative databases have been reviewed. Despite the differences in methodological strength, all can be used to evaluate the quality of these data sources. In the following section, common abnormalities in health care administrative databases that directly involve in the overall accuracy of their information will be discussed.



## 2.5 Anomalies in Administrative Databases

As administrative databases are becoming important sources of health care data, the anomalies within these databases are of increasing concern. In this section, we strive to address such existing anomalies by spacious review of literature so that, administrative databases can be properly used for research purposes despite their limitations. This would help one to understand the element structures and processes that influence and generate data in health care databases.

It can also be used as a framework to identify and discover anomalies within the RAMQ administrative databases. More implementations can be carried out to refine the framework and to develop a new physical scheme. This requires further assessment of the actual data through investigations and validation studies to examine plausibility, technical feasibility, completeness, and usefulness of the RAMQ databases.

Since *process-*, *element-*, and *interpretation-attributes* produce and influence anomalies, certain types of attributes may serve to predict specific types of anomalies. Continued research in the refinement of the framework categories and the development of a more comprehensive catalog of anomalies could then serve as a predictive tool for identifying anomalies and potential anomalies in a variety of health care databases.

A description of the types of anomalies is thus detailed along with descriptions of each category. The descriptions start with discussing errors and the occurrence of errors in the acquisition process, then the transformation process, and finally the analysis process. All missing anomalies in each process type will be considered and finally all variation anomalies in each process type will be discussed.

## **2.5.1 Error**

Although errors may easily occur as random errors, they may also occur as the misapplication of the rules designed to ensure data integrity. Examples are selecting the wrong value from a list of allowable values for a field, selecting the incorrect value to use in a given situation, or violating the rules for ensuring data consistency.

### **2.5.1.1 Acquisition Errors**

#### **2.5.1.1.1 Data Entry**

This includes such errors as transposition errors of numbers, misspellings, typographical errors, etc. These are more or less the random errors that may occur in collecting and recording data [Iezzoni LI, 1990b; Potvin L and Champagne F, 1986].

These errors are non-systematic and so are not included in the scheme herein presented.

#### **2.5.1.1.2 Invalid Value**

This may include a value that is out of range, lacks the correct number of significant digits for a valid identifier (such as for a social security number) or a value that does not match the data type (such as entering a character where a numeric should be entered) etc [Calle JE, 2000; Wolff N and Helminiak TW, 1996]. Some of these may seem to be random errors such as lacking the correct number of significant digits or entering characters where numbers should be. However, investigation into databases has shown that in fields where values were required, occasionally people purposefully entered incorrect entries simply to satisfy the demands of the program.

Unfortunately, separating whether errors are randomly or systematically occurring may be difficult in some cases. It may seem that invalid values should be eliminated through programming the criteria for validation into the interface or database. However, this is not always the case perhaps due to attempting to provide flexibility for allowing occasionally seemingly invalid values that are on occasion valid, or for allowing for differences in the expression of medical events, etc.

#### **2.5.1.1.3 Misrepresentation**

These are values that are valid, but are the wrong value for the situation. For example, as mentioned earlier, systems that restrict the recording of an observation to a coded list may risk misrepresentation if the correct code cannot be found such that a code that seems closest to the correct code is entered, yet it does not accurately describe the observation [Martin-Baranera M, 1995].

#### **2.5.1.1.4 Consistency**

These errors include the violation of some criteria that exists for a set of elements usually based on some condition or set of conditions [Martin-Baranera M, 1995]. For example, a patient is recorded as being pregnant while the sex of the patient is recorded as male. These types of errors often occur systematically when default values are used and the user doesn't change the default. Another example is if the same information can be recorded into more than one location and there is a difference in agreement between the two recordings.

## **2.5.1.2 Transformation Errors**

### **2.5.1.2.1 Transformation Algorithms**

Transformation errors may occur when transformation algorithms are applied to data in an attempt to conform data to standards and formats different than the original format and values [Wolff N and Helminiak TW, 1996]. The result could be one of four outcomes: valid data becomes invalid; invalid data stays invalid; invalid data becomes valid; valid data stays valid. The real problem may be determining which one of these outcomes has occurred. If this can not be determined, the reliability of the data is compromised.

Data may also become inconsistent or may no longer accurately represent the original situation. For example, attempting to conform multiple code sets into one standard code set may result in forcing multiple codes into a single standard code, or in a situation where a standard code has no equivalent in a source code set. In the first situation, a value may be over-represented whereas in the second situation a value may be under-represented.

### **2.5.1.3 Analysis Errors**

#### **2.5.1.3.1 Wrong Assumptions**

Analysis errors, in terms of the data, may result from misinterpretation due to assumptions made about the data that are not correct. For example: the assumption that a value is an error when it is not an error, or the assumption that a value is correct when it is not correct. Assumptions may also be made about the representation, consistency, and reliability of the data.

## **2.5.2 Missing Data**

### **2.5.2.1 Acquisition Missing Data**

There are several causes or reasons for missing data [Chowdhury S, 1991]. It is important to note that many of the reasons are appropriate, such as absence of data may indicate a normal value or recording a value would not be applicable to the situation. A breakdown of possible reasons, observed by the author and collected or determined from research is discussed in the following paragraphs. Although time is a consideration for all potential anomalies, it seems that it may play a significant part in promoting missing data. The less time a clinician has to collect and record data, the greater potential there is for these data to be missing [Chowdhury S, 1991].

**2.5.2.1.1 Not Applicable**

Entering a value into a specified field is necessary for the particular situation. For instance, if a patient were not transported to the hospital, then fields requesting data about the transport would not be applicable.

**2.5.2.1.2 Unknown**

The value was not known at the point when the data were being entered into the database. It could be the data were never known or lost at some point in the acquisition process.

**2.5.2.1.3 Not Measured**

The data element was never obtained. If the person gathering the original data is unaware there is a need for a particular data element, then it may be less likely the data will be obtained in the first place [Sackett DL, 1979].

**2.5.2.1.4 Not Noteworthy**

The person recording the data did feel the event or observation was noteworthy enough to record.

**2.5.2.1.5 Not Necessary**

The person believes the most important data are already present and so other data could be omitted [Chowdhury S, 1991; Sackett DL, 1979].

#### **2.5.2.1.6 Not Recorded**

The data were overlooked or forgotten when entered into the database [Chowdhury S, 1991]. In addition, some information may be sensitive in such way that the patient or clinician does not wish the information to be recorded [Harris HE, 1997].

#### **2.5.2.1.7 Mis-recorded**

The value was entered into the wrong or another place. In a systematic way this may occur when users are not well trained or software allows entering the same observation in more than one place. A prototype is software that provides pre-defined lists from which users' record observations may also provide another free-text field in which users may enter observations not on the list. It has been previously shown that some users used the other free-text field to record items regardless of whether they were on the pre-defined list or not.

#### **2.5.2.1.8 Lost Data**

The data were lost at some point in the acquisition phase or while in the database. This could be from system crashes, program bugs, or data maintenance errors resulting in the inadvertent deletion of data. This may sometimes occur unexpectedly during upgrades [Chowdhury S, 1991].



#### **2.5.2.1.9 Normal Values**

Values representing a normal situation or condition are not recorded [Sackett DL, 1979].

### **2.5.2.2 Transformation: Missing Data**

Missing data for transformation focuses on transforming only subsets of data. Some of this transformation is for the purpose of “*cleaning*” the data in an attempt to make the data more useful for a specific purpose. Some examples include:

#### **2.5.2.2.1 Exclusion of Rows and /or Records**

Selecting only a subset of rows for transformation based on some criteria (e.g. all Medicare patients for inclusion on Medicare forms).

#### **2.5.2.2.2 Exclusion of Fields**

Not including fields in a transformation based on some criteria. For example, excluding fields with a high amount of missing values, or assuming fields are not important or necessary for inclusion in the new data set.

#### **2.5.2.2.3 Exclusion of Values**

This is the exclusion of a value in a field based on some transformation rule. A field that allows characters may have the values containing characters removed during transformation so the remaining numeric values can be stored in a numeric type field. For example, in a pre-hospital care database diastolic blood pressure allowed the user to type in a numeric value or the letter “P” to designate that the blood pressure was palpated. If someone wanted to analyze the diastolic pressure as a numeric value, the field would need to be converted to a numeric type and the palpation designation would need to be recorded some other way. Another example can often be found in research databases where users enter values such as 99:99 for an unknown time which is also an invalid time, but these databases are constructed to hold these values because this method of coding is a standard. In addition, some values may be excluded for confidentiality reasons such as values that can be used to identify a patient as in social security number, or location of where a patient injury took place as in an automobile accident. These may occur as part of the “data cleansing” process in a data warehouse project.

#### **2.5.2.3 Analysis of Missing Data**

Anomalies caused by missing data in the analysis phase is often a function of the question or questions being answered and the data needed to answer those

questions such as only studying diabetic patients. In this way, there are various reasons why only subsets of the data may be used.

#### **2.5.2.3.1 Eliminating Rows/Records**

For example, eliminating rows with missing data, or because they do not fit some criteria. Analysis may also attempt to account for data that is assumed to be in error by deleting those rows containing values thought to contain bad or unreliable data or outliers.

#### **2.5.2.3.2 Eliminating Fields**

All fields may not be necessary for analysis, and in this way fields may be eliminated from the final analysis. A concern is if remaining fields were dependent on the field that was eliminated [Connell FA, 1987]. For example, there may be several fields that determine a patient is a diabetic patient such as diagnosis codes, lab results, and pharmaceuticals. Eliminating one of those fields may make it difficult to determine which patients are in fact diabetic.

#### **2.5.2.3.3 Eliminating Values**

Analysis may also attempt to account for data that are assumed to be in error or which may cause anomalies by deleting or excluding those values thought to be in error such as outliers. After the initial framework was developed, it became apparent that although analysts were involved in the processes that eliminated rows, fields, and

values they were still performing transformation functions. This was one of the reasons the initial framework was later revised.

## **2.5.3 Variation**

### **2.5.3.1 Acquisition Variation**

Variation in the acquisition process may exist at many levels from methods used to collect data to the structure of the data themselves. Variation may also exist in the events and decisions used to determine how patients are treated. Although these events affect the data, it is not the purpose of this paper to assess these events for their influence on variation in the data. The purpose hereby is help individuals using health care databases to know that any difference they may find in the data is real and not a result of some error or variations in the data themselves.

There may be many ways to categorize variation. A list of variations is included in the sections that follow. Although it may not be exhaustive, the list seems to represent variations that may be of concern. It may be important to note that variation usually only exists when comparing more than one database or data set. However, there are instances where variation may exist within a single database. Variation within a database may be due to a normal progression of change to reflect changes in the organization, practices or guidelines. Variation within a database may

also reflect the inconsistent application of rules to the data, which may result in compromising the data integrity [Brackett MH, 1996].

#### **2.5.3.1.1 Structural**

Structural variations are differences in how data are structured when entered into permanent storage. For example, this may include differences between structured entry, where data are broken down into defined fields, and free-text entry, where those recording the information are free to select their own values and/or define their own structure. This includes how dates are formatted, as well as numeric fields, etc. Differences in how data are organized and stored over time are also included. For example, systems that keep audit trails, or addenda for recording changes to records, may be keeping important information in separate files or tables.

Some of these data may be redundant and some may not. A system that is structured such that it is possible to store a single data value in more than one location, field or form, may also promote anomalies when attempting to analyze these data. A final example for structural variation is the level of detail or granularity recorded about an item.

#### **2.5.3.1.2 Definition**

Definition variation really includes differences in what the definition of a certain element is assumed to be by those recording the data. These definitions are more commonly referred as operational definitions. It is important to note how a software vendor defines what data should be entered into any particular field may not be what is in fact entered into that field. Operational definitions may also differ between individuals collecting data. For example, different individuals may have different criteria for deciding what severity level to assign to a patient.

#### **2.5.3.1.3 Method / Measurement**

Method variation covers a wide area because the methods, by which data are collected, recorded, edited and stored, may vary greatly. These differences may include human versus machine collection of data, or differences in how measurements are made. For example, if pulse is always counted for 6 seconds then multiplied by 10, the resulting value will always be a factor of 10 creating a stair step graph of pulse rates rather than a smoother line more reflective of actual values.

#### **2.5.3.1.4 Constraints / Rules**

Constraint variations can best be defined as those rules that are applied to data either through human enforcement or software enforcement in order to restrict or constrain what values may be entered into a field. This may include criteria used to

decide what values to enter such as standard coding practices, to relationships which are enforced within the data, to set default values, etc. Limits or ranges for certain data values would be an example of a constraint. The set of codes from which an individual may select would be another. Default values may be of special concern since it may be quite easy for someone to accept a default value. In this case knowing what fields have defaults, what the default values are, and then knowing how often the default occurs may be very important to analysts.

#### **2.5.3.1.5 Temporal / Historical**

Temporal variation really refers to the changes that may occur within a single database or system over time. Any change causes a difference between the data collected, before and after the change occurs. For example, the RAMQ database, with compiled data from health institutes across Quebec, had previously used codes that no longer existed and were now invalid when compared with the new set of codes (e.g., emergency visits before 1991).

#### **2.5.3.1.6 Record Keeping Practices**

Some record keeping practices may affect how the data can be used during analysis. For instance, consider a situation where ICD-9-CM codes for a number of non-billable diagnostic procedures, performed at hospitals, are not provided.

#### **2.5.3.1.7 Purpose**

The reason for which data were originally collected gives one an understanding of the type of data elements likely to exist and perhaps alerts one to the limitations on the values recorded [Huff ED, 1994; Potvin L and Champagne F, 1986]. For example, claims data may contain information on dispensed medications, however, they contain only those medications that are being reimbursed (NOT over-the-counter drugs or samples). Understanding that a claims database supports the reimbursement process may bring to light types of data that may or may not exist in the database. In addition, data in health care databases have not been collected with a specific research purpose, and this will affect the usefulness of these data in research [Sorensen HT, 1996].

### **2.5.3.2 Transformation Variation**

#### **2.5.3.2.1 How Data Are Transformed**

Variation occurring during the transformation process may be produced by the differences in how data are transformed. Differences include: changes which may occur as a result of altering the structure or granularity of data, or ways in which differences are reconciled such as merging different code sets or coding systems



[Connell FA, 1987]. These errors may occur as part of the “data cleansing” process as in a data warehouse project.

### **2.5.3.3 Analysis Variation**

#### **2.5.3.3.1 How Data Are Used**

Variation in analysis probably centers on how the data are used and what questions are being asked of the data. The proposed framework does not expand on this, as the variations can be many, though it may be important to note differences in what the data were collected for and how they are being used.

#### **2.5.3.3.2 Assumptions**

Since there may be assumptions being made about the data in any analysis effort, it may be worth recognizing and stating them. For example, assuming that all medications are entered into the system, when in fact only billable medications are entered.

**Table 2.6:** Summary of Abnormalities found in Administrative databases.

	PROCESS		
	Acquisition	Transformation (zero to many)	Analysis (zero to many)
	Potential Causes of Anomalies		
<i>Errors</i>			
	Data Entry Invalid Value Misrepresentation Consistency Other	Transformation Algorithm Other	Wrong Assumption Other
<i>Missing</i>			
	Not Applicable Unknown Not Measured Not Noteworthy Not Recorded Mis-Recorded Lost Data Assumed Normal Values Other	Exclusion of Rows/Recorded Exclusion of Fields Exclusion of Values Other	Eliminating Rows/Records Eliminating Fields Eliminating Values Other
<i>Variation</i>			
	Structural Definition Method / Measurement Constraints/Rules Temporal/Historical Record Keeping Practices Business Practices Purpose Other	How Data Are Transformed Other	How Data Are Used Assumptions Other

### 2.5.4 Assessing a Condition in Administrative Databases Using a Gold Standard Test

A condition can readily be detected using a 'gold standard' diagnostic test. However, applying the same diagnostic test to individual records obtained from a database would not result in the same estimation.

Claims databases are frequently used to estimate the incidence and the prevalence of a certain medical condition, as well as the number of services (e.g., hospitalization) provided to a target population. However, there are some limitations associated with the use of administrative databases. The potential errors and drawbacks in the use of administrative databases are previously noted. In this part, a practical example [Couris CM, 2002] will be given along with a method of correction to accurately assess a condition in a claims database.

Usually, records in claims databases include mandatory information on the patients, their diseases, and the procedures performed [Connell FA, 1987]. A prototype of using claims databases could be to provide national estimates of the number of hospitalized incident cancer cases. Nevertheless, false incident cancers due to a lack of specificity and unrecorded cancers stemming from a lack of sensitivity create a bias in the estimation of the number of incident cancers [Dussaucy A, 1994; Lombrail P, 1994; Greenland S and Robins JM, 1985]. As a result, *claims databases are an imperfect means* of providing estimates about hospital activity related to

incident cancer cases. As this information might play a major role in health care settings and decision-making [Roos NP, 1996; Wray NP, 1995], these estimates are important and one should ascertain of their accuracy. In case of uncertainty, a method should be implemented to correct the estimates yielding true parameters.

Some studies have evaluated the reliability of hospitals' claims databases in identifying incident cancers [Cooper GS, ; Solin LJ, 1994]. When claims databases were evaluated for the detection of incident breast cancer, using the medical records as the standard, the most frequent cause of incorrect identification was an incorrect diagnosis codes of breast cancer from a negative biopsy [Solin LJ, 1994]. In addition to limits arising from miscoding for medical information, the main difficulty was to differentiate incident cases from prevalent cases. Thus, algorithms combining diagnosis and procedure codes were used [Cooper GS, ; Solin LJ, 1994]. Finally, the fact that claims databases were originally intended for reimbursement rather than to provide information on patients' health status or medical treatment might also introduce bias in the detection of hospitalized incident cancer cases. A diagnosis may have been present but unrecorded on the computerized abstract either because there were more serious conditions present or because the condition was deemed as not contributing to the care provided to the patient. However, since cancer is a serious disease, this limit should be minor.

The goal here is to provide an analytical solution to correct estimates of the number of incident cases given by claims databases (with credibility intervals). This

method, proposed by Couris et al. [Couris CM, 2002], is based on a two-phase study design using an external data set as a source of sensitivity and specificity estimates. It was validated by a simulation and applied to correct the estimate of the number of incident breast cancers hospitalized in a large group of French hospitals.

Using a Bayesian approach, this study provided a corrected estimate and, mostly, a credibility interval of the number of women hospitalized with an incident breast cancer at the Hospices Civils de Lyon in 1997. The correction method took into account the measurement errors linked to the use of the French claims database. Nevertheless, it might suffer from the constraints due to the use of a beta distribution to describe the knowledge of sensitivity and specificity in prior information. However, this method might converge with more accurate estimations of the sensitivity and the specificity.

# **Objective & Hypothesis** **3**

---

## **3.1 Problem**

The management of health care system nowadays places major emphasis on significantly reducing cost and resource utilization. Therefore, to reduce costs, one must identify various characteristics that might influence resource utilization. On average, one of the most important factors contributing to the total cost in a health care system is in-hospital morbidities, including the patient's length of stay (LOS) and number of hospitalizations. Screening the trends of hospitalization has a major priority on finding appropriate ways for efficient use of resources. However, easy access to this useful information is not timely available and is always provided when permission granted. Consequently, there would be *enormous value in finding reliable ways to estimate episodes of hospitalizations in a routine and affordable manner.*

Medical records are frequently considered the gold standard data source to measure the quality-of-care indicators because they contain detailed clinical information [Steinwachs DM, 1998; Fowles JB, 1997]. However, high costs and restricted access to medical records limit widespread quality measurements using this data source. Instead, claims databases are an alternative data source.

Today, there is a growing use of administrative databases to report the health outcomes and to estimate quality-of-care indicators, including resource utilization and performance indexes, in a group of individuals and patients. Indeed, the development of existing information systems to readily capture data for epidemiologic and outcomes studies has become an interesting area of research. This increase in frequency may be due, in part, to advances in computer technology and the validity and widening scope of the database methodology. It also continues to be an ongoing challenge for program successes.

While the trials and tribulations of data management systems today fail to provide all of the desired research solutions, small advances can be made to provide key information. This can be, for example, *creating standard extraction tools and algorithms to capture vital information on surrogate measures of quality indicators from administrative databases so that they can be used in epidemiologic, economic and outcomes studies.*

### 3.2 Hypothesis and Objective

The capacity of the RAMQ medical claims data for estimating in-hospital morbidities is not being fully recognized. Furthermore, one must identify the types of information that, if available, would increase the potential for acquiring quality

information regarding certain quality-of-care indicators. Given the wide availability and relative cost advantage of the RAMQ claims data, it is worthwhile to understand whether they can be used to measure inhospital morbidities, including length of stay and readmissions.

Therefore, the main goals of this thesis dissertation are:

- To develop an algorithm to precisely capture episodes of hospitalizations, including admission and discharge dates, and to accurately estimate the length of inhospital stays in a target population (COPD patients), based on existing information in the RAMQ medical services database (training dataset).
- To validate this algorithm through comparison of the final results to those obtained from a gold standard data source—MED-ECHO database, using deterministic linkage method. This would ensure that the desired outcome indicators are appropriately estimated and pertain to those measured in the gold standard, which would establish the potential of the RAMQ data to measure inhospital morbidities.
- To illustrate that the results are consistent and reproducible when used in other medical conditions using an external validation technique (validation dataset). This will test the reliability of the algorithm.



Confirming a high level of accuracy and reliability of the data will increase the confidence of the conclusions of studies in which RAMQ is the primary source of data and the generalizability of administrative database methodology.

The use of this method lends itself readily to analyses, providing extremely large data sets, at very little cost. These advantages have already been discussed in the previously published literature [Roos LL and Brazauskas R, 1990; Roos NP, 1988; Roos LLJr, 1987; Potvin L and Champagne F, 1986].

# ***Methodology*** **4**

---

## **4.1 Introduction**

The use of analytical approaches in both fields of clinical epidemiology and outcome research has partly evolved as a result of advances in biostatistical techniques, sophisticated software, and computer literacy. *Analytic epidemiology*, which is the study of determinants and risk of diseases through analytical designs and techniques, offers the unique opportunity to carefully investigate the outcomes of clinical practices and pharmacotherapy in human subjects.

Epidemiologic designs and techniques are often used *to develop statistical tools and algorithms to better describe and in some cases predict the distribution of diseases, outcomes, and their influential determinants* resulted from various health care procedures. Over the last few decades great strides have been made in our understanding of these processes particularly those governing the final endpoints in clinical practice, ***Outcomes Research***.

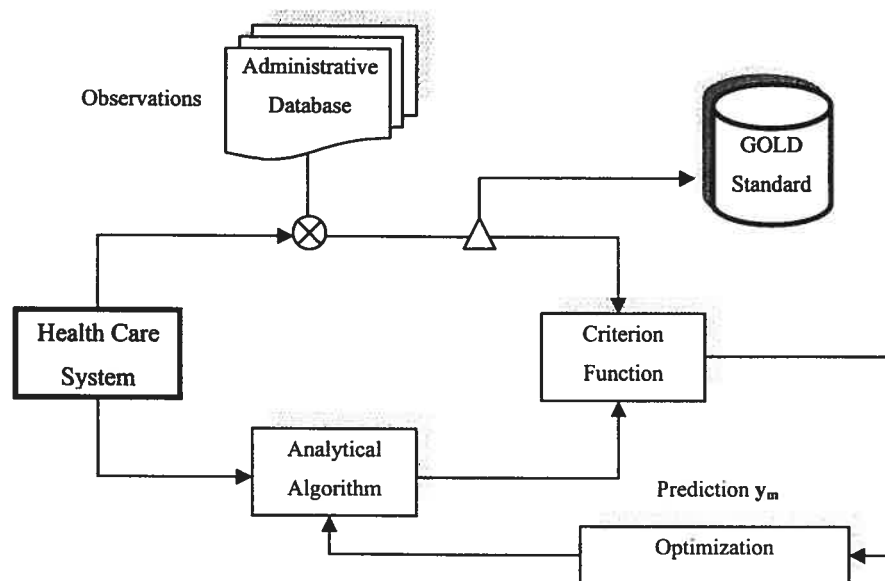
This growth has come with the advent of ***computerized databases*** in health care systems and appropriate techniques required to analyze these data. In fact, the use of computer technology and health care databases in clinical epidemiology and

outcomes research has recently allowed researchers to investigate problems that were once described as impossible or perplexing tasks, particularly in a large population. Outcomes research has thus graduated from being an essentially descriptive subject to one of vital practical interest to all those involved in the design, evaluation, and administration of health care systems. Overall, the ability to study final endpoints of clinical practices and interventions (including outcome indicators) with *speed*, *reliability*, *accuracy*, and a *reasonable budget* has been a great benefit for the advancement of scientific research in the field of analytic epidemiology and outcomes research.

Broadly then, the purposes of outcomes research are, on one hand, to *reduce and simplify data to a number of meaningful parameters* to be used in the current health care settings and on the other hand, to *evaluate the applicability of the results* for the future clinical practices which would be too costly and time-consuming to complete.

Measuring outcome indicators in a scientific manner, however, requires certain tools and algorithms, that allow researchers to extract, collect, and compare *valid* and *reliable* information. *Valid* refers to the concept that the results pertain directly to the desired attribute or characteristic being measured; whereas, *reliable* indicates that the results are consistent and reproducible [Testa MA and Simonson DC, 1996]. In the majority of outcome studies, as previously noted, the observational data are gathered from health care administrative databases. However, the quality of information in

these databases should be assessed through comparison to a criterion function (**Gold Standard**), which represents the real system. Figure 4.1 represents the functional diagram of a data validation process through stages of data sources, the analytical algorithm, the choice of criterion function (gold standard), and the optimization step(s). These four stages, including a specific analytical algorithm developed to meet the objective of this dissertation will be in detail explained in the following sections.



**Figure 4.1** – Four stages in a database validation study using linkage method.

## 4.2 Data Sources

Data are required to measure the quality-of-care indicators. Data collection can be performed by directly observing the practice of care providers or by studying the documentation and other records that are produced as health care is delivered. The best choice of a data source depends on available data content, accuracy, ease of use, cost, and the purpose of the quality measurement activity. In our study, two data sources for the measurement of hospitalization episodes were compared: hospital discharge summary (MED-ECHO) as the gold standard and medical claims data (RAMQ). During the years that were included for study, enrollees of the Quebec universal health insurance plan (more than 95% of residents in 1996) had the following services insured: in- and out-patient medical services, medications for persons over the age of 65 years, and hospitalizations. Records on hospitalizations (Quebec hospital discharge summary, MED-ECHO) are collected by the Ministry of Health and other records were compiled by the Quebec medical insurance board (also known as RAMQ). The electronic databases maintained by the RAMQ include: 1) medical registrant data; 2) physician claims data; and 3) medications dispensed by a community pharmacy to insured persons.

### 4.2.1 MED-ECHO

The hospital discharge database, referred to as MED-ECHO, is maintained by the Quebec Ministry of Health and collects information on all hospitalizations in acute care institutions within the province of Quebec. A pilot phase began in 1976 and the system has been fully operational since 1980. For each hospitalization, a discharge summary form (AH-101P) is completed by a local medical archivist. The data are collected by fiscal year, April 1<sup>st</sup> to March 31<sup>st</sup>, and does not include services provided in the emergency room or in out-patient clinics—an important drawback using MED-ECHO database. The same patient may have multiple hospitalizations and data entries in the same year.

The MED-ECHO database respects patient confidentiality by providing anonymous records which were examined for this thesis dissertation. All information regarding the MED-ECHO database are provided when a formal request is made to “La commission de l'accès à l'information du Québec” (CAI). Receiving data from MED-ECHO is a time-consuming process which can take several months. Each year there are about 1,000,000 entries into the MED-ECHO database. Each entry includes patient age, gender, principal diagnosis and the possibility of up to 15 secondary diagnoses, as well as medical procedures at different occasions.

Diagnoses are coded according to the ICD-9 [Practice Management Information Corporation, 2001], and treatments are coded according to the Canadian

Classification of Diagnostic, Therapeutic and Surgical Procedures [Statistics Canada, 1986]. For this study, a subset of data was obtained, including the patient's Medicare number, type of hospital, dates of admission and discharge, length of stay (LOS), principal diagnosis, and up to 15 secondary diagnoses, up to nine treatment codes, type of destination after discharge (home, dead, other acute care, rehabilitation, chronic care, and "others").

A number of studies have examined the validity of the MED-ECHO database [Levy AR, 1999; Levy AR, 1995; Mayo NE, 1994; Delfino RJ, 1993; Mayo NE, 1993]. In general, the coding was found to be very reliable for diagnoses where there are clear diagnostic criteria (e.g., hip fracture) and less reliable where the diagnoses were more ambiguous (e.g., stroke). For hip fracture, the accuracy of discharge data was 99% (198 out of 200 cases of hip fracture were correctly identified) at the level of the 3-digit ICD-9 code [Levy AR, 1995]. In a validation study of 679 records with respiratory diagnoses, agreement between the medical chart and the database ranged between 78% and 86% [Delfino RJ, 1993]. For conditions such as stroke that are more difficult to diagnose, the agreement between two neurologists on the diagnosis of stroke (based on chart data) and the database record for 96 medical charts was between 70% and 80% [Mayo NE, 1993]. In another study, the reliability of records with myocardial infarction (MI) on the MED-ECHO database have been evaluated to be very high with the positive predictive value of 96% [Levy AR, 1999].

### 4.2.2 RAMQ

All Quebec residents are covered for health services under the Provincial Healthcare Fund administered by the Quebec medical board (RAMQ), regardless of where care is obtained. Participation in the plan is virtually complete because residents are not required to pay any premiums. The enrolled beneficiary population closely matches the provincial population as determined by the census. The fund covers the costs of prescription drugs (only for elderly, welfare recipients, and individuals who do not access to private insurance), physician visits, and other medical services offered in private clinics or hospitals. However, a \$2 copayment per prescription (maximum of \$100/year) was required for elderly after its introduction in 1996. Low-income elderly patients (representing 5.7% of all elderly patients), as well as welfare recipients, had no copayments.

Beginning August 1, 1996, all elderly patients were required to pay a 25% coinsurance fee for prescription drugs. Annual ceilings for this copayment (\$200, \$500 or \$750, depending on personal income) were also introduced. As of January 1, 1997, an annual deductible was added to the program. The deductible varied from zero to \$350 depending on personal income. Starting on July 1, 1997, the coinsurance and the deductible, which until then had been divided quarterly, were prorated monthly to reduce the amount of any one payment. As a result, the maximum amount that a person pays for prescription drugs per month now ranges from \$16.66 (\$200 ceiling) to \$71.42 (\$857 ceiling) [RAMQ Official website, The Public Plan].



RAMQ claims data are generated for billing purposes as a result of a patient's encounter with the Quebec health care system, including outpatient care, hospital care, and filled prescriptions. For reimbursement, all health care providers (including pharmacists, physicians, and hospitals) must submit claims to the provincial government. All claims are captured by the RAMQ databases which contain information on patient, as well as medical and pharmaceutical services they have received. Enrollment data are also maintained electronically by health plans to identify the people who are eligible for drug insurance coverage. Together, claims and enrollment files contain information on demographics, diagnoses, delivered services, and dispensed prescriptions. Details about these files will be discussed in the next section.

Although RAMQ claims data have less clinical information than hospital discharge abstracts, they are widely available and relatively inexpensive to obtain and analyze. RAMQ Claims can also be feasibly de-identified (encrypted) which minimizes concerns about privacy and confidentiality of the records. *Since RAMQ claims data are routinely collected and computerized, it could be potentially used to estimate hospital morbidity indicators over time to identify trends and progress in quality.* A large number of cases generally contained in RAMQ claims files also permit multiple comparisons, the testing of hypotheses about population subgroups, and comparisons across statistical models.

Since RAMQ claims data can be easily accessed, i.e., more time-efficiently and less costly, compared to other data sources (i.e., MED-ECHO database), *they have the potential to contribute to the knowledge base about hospitalizations for various medical conditions*. However, any data source used in research should be evaluated with regard to two criteria: **availability** and **accuracy**. Evaluating the *availability* of a data source means understanding who and what activities are included in the data source and exactly what type of information the data contain. *Accuracy* addresses whether the data source can generate *reliable* answers to the quality question at hand. These concepts will be later discussed comprehensively in this chapter.

#### 4.2.2.1 Medicare Registrant Database (Enrollment)

The medicare registrant database contains the Quebec health insurance plan unique identification number, name, address, Centre Local de Services Communautaires (CSLC) and forward sortation area (first three digits of the postal code) of residents, gender, date of birth (if permission granted by the CAI), and date of death (if deceased) for all Quebec residents who are covered by the provincial health insurance program (approximately 97.7% of Quebecers in 1991 and almost 100% at present). In order to safeguard confidentiality, the name and address are not supplied; the unique identification number is scrambled. Age is reported in years based on the age on July 1<sup>st</sup> of each year and is therefore subject to a potential

misclassification of up to six months. This misclassification is non-differential and generally has only a very small impact on the results.

#### 4.2.2.2 Physician Claims Database

The physician claims database contains information on the medical services provided by physicians or other health care providers on a fee-for-service basis in Quebec. In 1989, 95% of all reimbursed medical services were estimated to be paid for fee-for-services. More recently it was indicated that more than 85% of family physicians and general practitioners across the province are paid in this manner [Blais R, 2003]. Thus, most physician visits are recorded and included the following information: the patient's medicare unique identification number, the physician's license number (scrambled) and specialty, the date of service delivery, a code for a medical service or procedure performed, the place of the service delivery (including hospital ward, intensive care unit, emergency room, outpatient clinic, private office), a diagnosis for the visit (coded according to the ICD-9; about 87% complete in 1989 [Régie de l'Assurance Maladie du Québec, 1994]), and the reimbursement for the visit.

Information in RAMQ physician claims data about patients' clinical conditions is most often in the form of *diagnostic codes* specified by the *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM). ICD-9-CM codes are used by healthcare providers to report their diagnostic assessments;

however, there is only one space allocated for registering the code, which is not mandatory, in the RAMQ claims.

ICD-9-CM contains more than 10,300 codes. The diagnostic codes of ICD-9-CM are organized within broad categories. Some of these categories represent various types of conditions (e.g., infectious and parasitic diseases, neoplasms), while others reflect anatomic locations (e.g., circulatory, digestive, respiratory systems) and one category is reserved for “symptoms, signs, and ill-defined conditions”. Three-, four-, and five-digit codes are listed, representing increasing levels of specificity. For example, the three-digit code 250 indicates diabetes mellitus, while the fourth digit specifies the manifestation (e.g., 250.5, diabetes with ophthalmic manifestations) and the fifth digit reflects the type (e. g., 250.52, diabetes with ophthalmic manifestations, adult-onset type). For some disease classifications, only four digits are specified [Practice Management Information Corporation, 2001].

Furthermore, information about performed *procedures* or *delivered services* are coded using the uniform billing codes in Quebec or the Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures (CCP) [Statistics Canada, 1986]. They may be transformed into the ICD-9-CM procedure codes. Inpatient hospital procedures must be reported using the ICD-9 Procedure Coding System (ICD-9-PCS), while a combination of CPT or CCP codes are required for physician services and other health care services.

#### **4.2.2.3 Physician Demographic Database**

The physician demographic database contains: the physician's license number (scrambled) and specialty of the physician who forwarded the claim to the RAMQ for reimbursement, gender, and year and medical school of graduation. These data are originated from the Quebec College of Physicians. The medical specialty is based on certification by the Federation of Medical Specialists of the Province of Quebec, and physicians have to successfully complete the Quebec specialty examinations before being certified. This means that some physicians with specialty training from other jurisdictions who have not taken or passed the Quebec specialty examinations are classified as general practitioners (GP).

#### **4.2.2.4 Prescription Claims Database**

The prescription claims database contains information on all dispensed prescriptions medications dispensed by community pharmacies to Quebec Medicare registrants 65 years of age and older, social welfare recipients, and other adherents (approximately 50% of the total population). The prescription claim record that is filled out by the pharmacist includes: the patient's unique identification number, the prescribing physician's license number, dispensation date, a DIN code for the medication that indicates the drug, manufacturer, dosage and mode of delivery,

quantity of the drug dispensed and the duration of prescription, whether the prescription is a refill, the form of the prescription (written or phoned in), as well as the reimbursement for the dispensed prescription. RAMQ prescription claims files that track dispensed medications generally use the Drug Identification Number (DIN), in which each human drug is assigned a unique identifier. Drugs dispensed to patients during stays in hospitals or public nursing homes, and over-the-counter drugs are not included in the database.

In general, the RAMQ database is thought to accurately reflect the dispensation of medications because pharmacists must submit a claim in order to receive payment (financial incentives) and routine monitoring and other control measures are used to limit over-reporting. The comprehensiveness and the accuracy of information recorded in this database are equivalent to or better than that reported for many registries that use primary collection methods. As previously stated, of 1,917,214 records of dispensed prescriptions, out of range or missing values in the fields for the individual identifiers, drug, quantity, date dispensed and duration ranged between 0 and 0.4% of records [Tamblyn R, 1995]. The accuracy of the information on medications in this database was also examined in 723 prescriptions filled by 306 elderly patients attending one internal medicine clinic. Of these dispensed prescriptions, 83% were filled by the patient and correctly identified the patient and the drug; in 89% of 5,999 records, the prescribing physician was correctly identified

[Tamblyn R, 1995]. The quantity and duration of the dispensed prescriptions were accurate in 69.1% and 72.1% of records, respectively.

## **4.3 Definitions**

### **4.3.1 Available Information on Claims Data**

Claims data are a by-product of reimbursing for health care services. Therefore, claims data generally include people who have health insurance, receive health care products and services, for which claims are being created.

Although the precise contents of claims database vary by health plan or insurer, RAMQ claims forms capture patient characteristics, provider identifiers, treatment and diagnostic information, pharmaceutical services, and payment information (Table 4.1).

Although RAMQ medical claims data contain all information regarding inpatient and outpatient encounters, they do not have direct information regarding hospitalizations—structural limitation, require specific algorithm to deduce these information. Furthermore, RAMQ data contained inhospital, outpatient, and prescription claims are limited to codes for diagnoses, services, and medications. These codes are frequently from, or can be linked to (e.g., procedural codes) standardized coding systems, facilitating quality measurements.

**Table 4.1:** List of Standard Data Elements in the RAMQ Medical Database.

Categories of Variables
<b>Patient Characteristics</b>
Patient Identifier
Date of Birth
Gender
Date of Health Insurance Coverage
<b>Provider Identifier</b>
Physician Identifier
Year of Graduation
Specialty
Hospital/Facility Identifier
<b>Diagnostic and Treatment Information</b>
Diagnostic Code
Date of Service
Place of Service
Procedures, Services or Supplies Codes
Code for Consultation
<b>Insurance/Payment Information</b>
Amount Paid
Total Charges

### 4.3.2 Accuracy and Agreement

For the purposes of this dissertation, agreement and accuracy will have to be defined. *Agreement* is the proportion of linkable records for a given variable in at least two databases that are the same or relate to the same outcome (for example,



diagnostic codes in two distinct databases that map to the same diagnosis or hospitalization episodes that are identified using the RAMQ claims database and those observed in MED-ECHO). For variables with clear definitions (e.g., patients' sex), which can be simply classified, agreement can be expressed mathematically as a percentage or a proportion. Where there is further definition or judgment required to assess an outcome variable (e.g., determining a diagnosis or assigning a procedure code), agreement and accuracy may be different. *Accuracy* is the proportion of linkable records for a given variable in at least two databases that relate to the same outcome *where the outcome can be judged true by standard clinical criteria* (if based on the clinical evidence they are found to be *true*). While ultimately it is most desirable to have accuracy information on all variables within a database, determining the accuracy for variables which may require judgment is limited by the availability of standard definitions or criteria for each variable. Thus, in the absence of these criteria, one may use the overall agreement rate as a surrogate measure of the variable's accuracy [Williams JI and Young W, 1996b].

## 4.4 The Logic of Comparison

Documenting the internal logic of a single database provides a limited approach to data quality; typically, only unobvious errors are revealed. For example, men should not be having hysterectomies; individuals should not have contact with the health care system after death; and so on. Such assessments might include:

- Diagnoses and procedures relevant to only one gender
- Diagnoses which should not occur after a particular surgery
- Events which should only occur in a particular age range
- Utilization which should not occur after death

Such a screen can highlight problems involving particular elements or particular individuals in a database. Explicit assessments then need to be made; if a man is noted as having a hysterectomy, which was coded incorrectly—gender or surgical procedure?

Data sources can be compared in several ways. One approach builds on the use of *individual health identification numbers* to link and compare information routinely collected from two data sources. A second approach uses *two or more sources with at least one source* involving considerable extra effort at data collection (prospective or retrospective).

Several studies have used survey or laboratory data, most have involved reabstraction of hospital charts for comparison with the original hospital discharge abstracts. Reabstraction studies examine how reliably the information from hospital charts is coded onto hospital discharge abstracts. Hospital charts are typically taken as the “*gold standard*” in reabstraction studies. In addition, very few studies have used multiple methods to establish the degree of agreement among more than two data sources [Muhajarine N, 1997]. Often, only those records with the condition of interest are taken from the first source, making overall consistency impossible to calculate from the published data.

In practice, defining a true gold standard involves considerable effort using the best available technology to try to establish the accuracy of a categorization. Prospective data collection may provide a gold standard, but such work is very expensive. Another broadly accepted method is the chart audit that has been used extensively in validation studies as the gold standard; however, it is also expensive and time-consuming as compared to electronic data comparison. Most studies take *one type of record as a gold standard* in data quality assessment. In our study, physician claims for episodes of hospitalizations were compared against an individual's hospital discharge summaries (MED-ECHO). The process of data collection in MED-ECHO is carried out independently which virtually covers all hospitalization records in the province of Quebec (“*universal information on hospital admissions*”); the validity of its diagnostic coding has also been established by several

studies [Pilote L, 2000; Levy AR, 1999; Choinière R, 1999; Levy AR, 1995; Mayo NE, 1994]. Thereby, the number of records with concordant information in the two data sources was matched, providing a basis for comparison with which physician claims' quality was measured. In the remaining of this chapter, steps required to develop an analytical algorithm to estimate inhospital morbidity indices, i.e., length of stay (LOS) and number of hospitalizations, will be explained. Several programming techniques and statistical parameters have been used to provide a flexible framework for parameter estimation and comparison of different algorithms. Furthermore, statistical parameters necessary to measure the level of accuracy and agreement of the developed algorithms will be discussed.

## 4.5 Algorithm

### 4.5.1 Introduction

Two studies were performed to assess the quality of the RAMQ medical services claims regarding in-hospital records, both of which involved administrative database linkage methodology. In the first study, we primarily developed an analytical algorithm to identify in-hospital morbidity indices, including length of stay and number of hospitalizations, from the RAMQ medical services claims. The **validity** (agreement) of the hospitalization records, derived from the RAMQ medical claims, was then assessed by comparing the parameter estimates to the corresponding discharge summary data provided by the MED-ECHO using a record-matching algorithm (MED-ECHO database was considered as the gold standard). Furthermore, the reliability of the study results was examined using diagnostic performance parameters and associated plots.

In the second study, the **reliability** and **robustness** (generalizability) of the developed algorithm in other clinical conditions, namely myocardial infarction and upper gastrointestinal bleeding, was examined using an external set of data provided by the RAMQ with the hospital discharge summary of MED-ECHO as the gold standard.

In the following section, the steps taken (according to their priority) to develop an analytical algorithm will be clearly outlined to permit the reproducibility of the analysis and model building. Moreover, the criteria and rationale for building procedures will be clearly stated.

#### **4.5.2 Study Sample**

This first study involved a secondary data analysis of a cohort previously constructed by the team of Pharmacoepidemiology and Pharmacoeconomics Unit of the CHUM–Hotel-Dieu Research Center, Montreal, Quebec [Blais L, 2004]. The constructed cohort was based on data selected from the administrative health databases of the Régie de l'assurance-maladie du Québec (RAMQ), the government body responsible for the administration of health programs available to residents of Quebec, a Canadian province of 7 million inhabitants. At the time of the study, between January 1, 1990 and June 30, 1996, the cost of prescription drugs for all residents over the age of 65 was covered by the RAMQ. In 1996, more than 95% of Quebec residents over 65 years old were RAMQ enrollees who benefited from the universal health program offered by the provincial government [Régie de l'Assurance Maladie du Québec, 1996]. Moreover, the RAMQ was also responsible for the

remuneration of physicians and hospitalization services for all residents of the province.

RAMQ administrative files were used to assemble a cohort of individuals over 65 years old (upon entry into the cohort) who had been diagnosed with Chronic Obstructive Pulmonary Disease (COPD) in Quebec during the study period (training dataset). Patients were selected upon having a claim with a diagnosis of COPD (ICD-9 codes indicating a diagnosis of COPD): non-specific bronchitis (490.X), chronic bronchitis (491.0, 491.1, 491.2, 491.8, 491.9), emphysema (492.0, 492.8), and chronic airway obstruction (496.9). Furthermore, selection was restricted to only those who had used bronchodilators chronically and had no previous diagnosis of asthma during the study period [Blais L, 2004]. Hospital discharge summary prepared by the MED-ECHO were used to identify patients admitted to a hospital with a diagnosis of COPD. Moreover, all other relevant records from these two data sources were extracted for the same period.

### **4.5.3 Development**

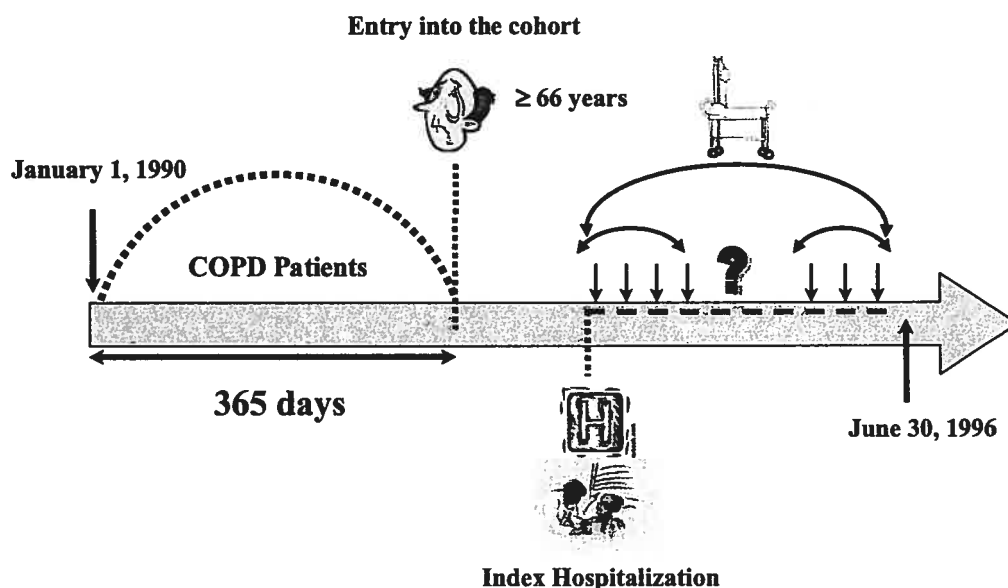
Using SAS computer software (SAS Institute Inc., Cary, NC, USA), Release 8.02, an analytical algorithm was developed to first identify inhospital records in the target population and then to estimate a set of defined inhospital morbidity indicators (inhospital length of stay and number of hospitalizations) using the RAMQ medical

services claims. This algorithm was then applied to all the separation episodes within the defined period. A separation episode is referred to an administrative process by which in-hospital completion of treatment and/or care are recorded; it is also known as discharge summary [Roos LL, 1996b].

The last stage of this project involved an independent record-matching algorithm by which the accuracy and validity of the parameter estimates using RAMQ medical services claims was compared to that of derived from the gold standard, MED-ECHO hospital discharge summary. The level of sophistication was based upon the unique health care identifier (patient's identification number), the information contained in the database, and the programming capacity.

Summarizing our methods, RAMQ medical services claims were used to first identify inpatient services and then to estimate length of stay (LOS) and number of hospitalizations. The validity of the parameter estimates was then assessed using hospital discharge summary database as the gold standard. Following is the description of steps taken for the implementation of the algorithm in order to correctly estimate episodes of hospitalizations. In addition, Figure 4.2 demonstrates the current situation in the RAMQ medical services database, as well as the performance of the developed algorithm.





**Figure 4.2 – Algorithm to estimate episodes of hospitalizations.**

The main objective of the algorithm was to correctly identify the episodes of hospitalizations when compared to the gold standard (MED-ECHO). In this example, it is not clear whether the interruption of the services has promptly occurred at the end of the first hospitalization, or it was simply the same hospitalization with no visit by a physician. In other words, it could be one hospitalization with a total length of stay of 10 days, or two distinct hospitalizations with the duration of 4 and 3 days, respectively.

- ⇒ **Step 1: Creation of an analytical file to estimate length of stay and number of hospitalizations in RAMQ** (This step was performed for the whole study period of 1990-1996).
- *Step 1a:* Using RAMQ medical services claims, an inpatient service finder was developed to create an analytical file with in-hospital claims.

This file contained all physician claims filed for inpatient services according to institutional codes subject to the hospital global billing;

- *Step 1b:* All physician claims including an inpatient code contained in the finder file were selected. Although the identification of inpatient codes provided essential information to establish the fact that a patient was certainly admitted to the hospital, we additionally needed an analytical algorithm to estimate the episodes of care;

- *Step 1c:* Claims were sorted by date so that they appeared in a chronological order; a single claim was retained, dropping multiple claims from the physician file. For the purpose of this analysis, multiple claims were defined as those with the same information on the following variables: beneficiary identification number, institutional code, and date of rendered services. This was necessary because only one claim per day was required to identify episodes of hospitalizations; yet a patient could have been seen by several physicians in a given day. However, for the purpose of matching on a clinical condition under study (e.g., COPD diagnosis in this data set), we retained available information on all claims, including beneficiary health care identifier, date of services, the procedure codes, and the ICD-9 diagnostic codes (if available);

○ *Step 1d*: All claims were followed until an interruption in the services (*gap*) was identified. This gap was assumed to be the result of a patient not being seen by a physician for at least one full day, while the patient was still hospitalized. The index date of the first hospital admission (*admission date*) was defined as the first in-hospital claim filed by the physician for each individual in a chronological order.

⇒ **Step 2: Introduction of a “GAP” to compensate service interruption in the RAMQ medical services claims.**

○ *Step 2a*: It is very likely that a patient does not receive any medical services for a certain period while being hospitalized. To correctly estimate the episodes of care, we assumed that there was a service interruption (*gap*) in which the patient was still hospitalized. Accordingly, several algorithms were built based on a predefined gap between 1 to 15 days;

○ *Step 2b*: If the period of service interruption was equal or smaller than the predefined gap, the total duration (i.e., the time between the first and last in-hospital claim including the gap) was considered as a unique hospitalization. The *discharge date* for any given hospitalization was defined as the last claim in a continuous manner, in

a period during which any service interruption was equal or smaller than the predefined gap;

o *Step 2c*: In order to obtain an optimal algorithm estimating LOS and number of hospitalizations, we modified the algorithm by adopting different intervals (gaps) between the services. As stated earlier, the gap could take a value between 1 to 15 days. This allowed for the comparison of estimated LOS and number of hospitalizations obtained from different algorithms to that of gold standard, which in turn would ensure the correct estimation of the hospitalization episodes;

o *Step 2d*: Duration of the continuous medical services provided to a patient is the inhospital length of stay and can be calculated according to the following equation:

$$\text{LOS} = [\text{discharge date} - \text{admission date}]$$

o *Step 2e*: After estimating the index hospitalization in the RAMQ database, the algorithm was run on the rest of the claims to find the next date in which another inpatient service claim was filed. This was the date immediately after the first discharge date in the longitudinal medical services claims. This date is considered as an initial date for a separate episode of care (hospitalization).

⇒ **Step 3: Creation of an analytical file for record-matching using hospitalization records in MED-ECHO.**

- *Step 3a:* All hospitalizations that occurred in acute-care hospital facilities from the MED-ECHO database were selected. We retained beneficiary health care identifiers, dates of admission and discharge, the principal ICD-9 diagnostic codes and up to 15 secondary diagnoses (if applicable);
- *Step 3b:* All hospitalizations reported during the study period were kept; thus those beyond the restricted period were removed. We further restricted our selection to only those hospitalizations that took place in a full fiscal year according to MED-ECHO (between April and March of each year);
- *Step 3c:* Duplicate records (less than 0.5%) and/or overlapping hospitalizations from the MED-ECHO database where two hospital stays shared similar admission and/or discharge dates were removed;
- *Step 3d:* An analytical record-matching algorithm for the linkage of hospital discharge summary (MED-ECHO) with information on hospitalizations derived from the RAMQ inpatient physician claims was developed. This procedure allowed the comparison of estimated

parameters (i.e., LOS and number of hospitalizations) with the gold standards.

⇒ **Step 4: “Linked” MED-ECHO–RAMQ records** (merging of the MED-ECHO hospital discharge summary with RAMQ derived in-hospital records).

- *Step 4a:* All hospitalizations with the admission date occurring within the first week of the study period in the RAMQ claims database were removed. Because it was likely that those hospitalizations were indeed the continuation of previous claims in the RAMQ;

- *Step 4b:* At the beneficiary level, using the MED-ECHO admission and discharge dates for each estimated hospitalization as anchors, all hospitalizations identified using the RAMQ claims database were linked to the appropriate hospitalization creating one analytical record per beneficiary per admission;

- *Step 4c:* For a linked record to be created, the date of the physician claim must be on (the exact time window) or within a day of the admission date (the expanded time window) of the hospitalization record in the gold standard data source (admission  $\pm$  1 day);

○ *Step 4d*: Only hospitalizations where the date of hospital admission in MED-ECHO with the estimated date of admission in RAMQ coincided were considered to be perfectly matched. This matching process did not depend upon the hospital LOS. According to the MED-ECHO database, over- and under-estimation in the LOS were identified.

○ *Step 4e*: Based on the presence or absence of hospitalizations in MED-ECHO, the  $2 \times 2$  contingency tables were constructed to calculate measures of agreement. The latter assumed the MED-ECHO hospital discharge summary to be more complete (gold standard) than the physician claims data.

#### 4.5.4 Statistical Analysis

For the validation purpose, we used a deterministic linkage method. Conveniently, the structure of RAMQ databases, having a unique identification number, permitted the examination of the completeness and accuracy of its data through comparison analysis to the gold standard.

Descriptive statistics were carried out to compare the estimated LOS and number of hospitalizations obtained from the RAMQ medical services database to those observed in the MED-ECHO file. The predictive performance of each algorithm (defined in the following sections) was measured using the observed and estimated parameters.

Furthermore, the accuracy of the RAMQ medical claims data concerning inpatient services was evaluated using the concepts of **diagnostic test** evaluation. Within this framework, MED-ECHO discharge summary data was considered to be the “gold standard” for defining the presence of a hospitalization, a diagnostic condition, or the performance of a particular medical procedure. Accordingly, Bayesian theorem was used to estimate measures of agreement of the algorithm identifying episodes of hospitalizations in the RAMQ medical services database. These measures include overall agreement (efficiency), sensitivity, specificity, and positive- and negative predictive values. Kappa statistics (Cohen’s kappa), as the



primary reliability measure, were also generated for each condition to measure agreement while controlling for chance agreement. Kappa statistics were then adjusted to take into account the perceived prevalence of the disorder and the relative frequency of positive and negative observations. If other statistic indices were used in our analyses they would be discussed when applicable.

#### **4.5.4.1 Descriptive Statistics**

The analyses were performed using SAS program (SAS Institute Inc., Cary, NC, USA) release 8.02, with days in the hospital or inpatient visits as the primary units of analyses.

Various descriptors of central tendency were examined, including arithmetic mean, geometric mean (the average of a log-transformed variable), mode, and median. In addition, the standard deviation and 95% confidence interval were calculated for all parameter estimates in an attempt to better define the central tendency. These variables were compared with graphs of the data to explore their usefulness in describing the central tendency, particularly when they were used for predictive performance analyses. Moreover, probability density functions (*pdf*) of different variables were also computed in order to better describe the '*typical*' value of the parameter. Proportions (%) are only used when describing the distribution of an event

that can arise only once at a consultation (e.g., age, gender, etc.) or to describe the distribution of events within a class of events.

Data sets with skewed distribution were subjected to logarithmic transformations to induce normality and in an attempt to symmetrize the distribution of variables to allow for parametric testing. Although this approach simplifies statistical analysis, it can alter the pattern of the distribution and decrease the identification of outliers. Accordingly, log transformations (natural logarithm) were performed and their mean values (geometric) calculated. Nonparametric or distribution-free tests such as Mann-Whitney-Wilcoxon or the median tests [Samuels ML and Witmer JA, 1999; Hogg RV and Tanis EA, 1988] were also performed, which are more appropriate for non-normally distributed data. These tests assume that the distributions differ only in their location, that is the shape and the variance of the distributions are the same. Comparison of parameter estimates was thus carried out using paired t-test for means and geometric means, whereas Wilcoxon signed rank test for medians.

#### **4.5.4.2 Predictive Performance**

There are several methods that can be used to evaluate the predictive performance of a new model or algorithm. One of these methods is to calculate the

percentage of prediction errors (%PE) and the percentage absolute prediction error (%APE) as the cumulative error of the stream of predictions (*percent deviation*) [Wu G, 1995]. These measures can be defined as follows:

$$\%PE = \frac{\text{Estimated value} - \text{Observed value}}{\text{Observed value}} \times 100\%$$

$$\%APE = \left| \frac{\text{Estimated value} - \text{Observed value}}{\text{Observed value}} \right| \times 100\%$$

Absolute predictive performances were also calculated according to the methods described by Sheiner and Beal [Wu G, 1995; Sheiner LB and Beal SL, 1981]. The absolute predictive performance includes: mean prediction error (ME) as *bias*, mean absolute prediction error (MAE) as *accuracy*, and root mean squared error (RMSE) as *precision*. Although some these measures were first introduced in the field of clinical pharmacology, they are presently being used in a variety of validation studies, including epidemiology, economics, psychology, and prediction forecasting [Engels JM and Diehr P, 2003; Dickerson RN, 2002; Johnson TS, 1999; Covey MK, 1999; Ettner SL and Notman EH, 1997; Walter SD and Cook RJ, 1991].

*Mean error* can be obtained by calculating the estimation error (*pe*), i.e.

$$pe = \text{Estimated value} - \text{Observed value}$$

then by calculating:

$$ME = \frac{1}{N} \sum_{i=1}^N pe_i$$

Accordingly, *mean absolute prediction error* (accuracy) can be defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |pe_i|$$

Another measure of absolute predictive performance is the *root mean squared error*, which is representative of the size of a typical error (precision) and is measured in the same unit as the data. It can be defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N pe_i^2}$$

The mean error (ME) and percentage prediction error (%PE) are *signed* measures of error which indicate whether the forecasts are *biased*—i.e., whether they tend to be disproportionately positive or negative. The mean absolute error and percentage absolute prediction error reflect the magnitude of disparity; they can be considered as summary measures of predictive accuracy. They have advantages over the previous statistics because using the absolute values keeps the magnitude of the errors visible. These measures are also similar in magnitude to, but slightly smaller than the RMSE; however, they do not square the prediction errors and, thus, are not overly sensitive to larger estimates (outliers). The root mean squared error is the most important measure of the predictive performance, which indicate the precision of the

estimates. It is made up of two separate components: error variance and bias (Mean Squared Error = Error Variance + Bias<sup>2</sup>). The RMSE penalizes outliers more because the difference term is squared. The predictive performance measures are usually presented as means [Sheiner LB and Beal SL, 1981], however, for in-depth comparison they are presented as mean plus 95% confidence intervals in this thesis.

#### **4.5.4.3 Measures and Levels of Agreement**

Many of the statistical indices generated in this project could have been produced by the SAS system. However, a spreadsheet has been used to provide comprehensive statistics for the assessment of diagnostic criteria and the reliability of the RAMQ medical claims data, where our investigations yielded data that could be summarized in a  $2 \times 2$  contingency table. As well as a wide range of indices of test or rater performance, confidence intervals for these quantities were also calculated by the spreadsheet [Mackinnon A, 2000]. An example of this spreadsheet using hypothetical data is given in Appendix 1.

In our study, investigation of the performance of diagnostic criteria and assessment of the data reliability in the RAMQ medical claims database gave rise to four-fold ( $2 \times 2$ ) contingency tables cross tabulating the presence or absence of an event in the MED-ECHO database against its putative presence or absence as

indicated by the test under study. Cells in the table correspond to true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of events, as shown in Figure 4.3.

**Figure 4.3 – Schematic overview of analysis of agreement in claims data.**

		<u>Diagnostic Test Results</u>		
		Yes	No	Total
<u>Criterion (Gold Standard)</u>	Yes	$n_{11}$ TP	$n_{12}$ FN	$N_{1+}$
	No	$n_{21}$ FP	$n_{22}$ TN	$N_{2+}$
		$N_{+1}$	$N_{+2}$	$N$

Findings on the data quality were summarized in terms of percentage overall agreement (efficiency), sensitivity, specificity, positive- and negative-predictive values of the RAMQ claims data because very few studies present any other measures. The table in Figure 4.3 is used to explain the analysis. These five measures of agreement and how they can be calculated from the table in Figure 4.3 are briefly described presently.

#### 4.5.4.3.1 Efficiency (Overall Agreement)

Overall agreement, also referred to as *efficiency*, is a statistical summary of concordance that ignores distinctions between positive and negative agreement (i.e., does not separately evaluate how closely the data sources agree about who is a “yes” and who is a “no”). With reference to the tables in Figure 4.3, the overall agreement rate is:

$$\text{Overall Agreement} = \frac{n_{11} + n_{22}}{N} = \frac{TP + TN}{N}$$

Since measures of overall agreement do not distinguish whether the disagreement stems from claims data underestimating (i.e., from poor positive agreement) or overestimating (i.e., from poor negative agreement) the number of patients who satisfy the eligibility criteria, measures of sensitivity and specificity were also used in this analysis.

#### 4.5.4.3.2 Sensitivity

In epidemiology, sensitivity (true positive rate) is one of the measures of the validity of a screening test; it is defined as the conditional probability that a diagnosis (or an event) within the specified group was coded on the RAMQ claim record given

that it was actually present in the gold standard, MED-ECHO. In other words, sensitivity is the probability of testing positive if the event is truly present. In this analysis, sensitivity evaluates how well one data source agrees with the other about whether an indicator's criteria for eligibility have been satisfied. Since the hospital discharge summary database (MED-ECHO) reveals the truth, the sensitivity was estimated as:

$$Se = \frac{n_{11}}{N_{1+}} = \frac{TP}{TP + FN}$$

When the hospital discharge summary is taken as the standard and indicates “yes” (cells **TP** and **FN**), the sensitivity of RAMQ claims data reports the probability that claims data will agree (cell **TP**). In other words, the sensitivity indicates how likely it is that the RAMQ claims will say “yes”, given the MED-ECHO data says “yes”. High rates of sensitivity indicate that a data source is not substantially underestimating the number of patients who satisfy the eligibility criteria relative to the other data source.

#### 4.5.4.3.3 Specificity

Specificity (true negative rate) measures how closely each data source agrees with the other on negative assessments. It is the conditional probability that a



diagnosis (or an event) was not present on the RAMQ claim record, given that it was not coded on the record from the gold standard, MED-ECHO. As with sensitivity, for the RAMQ claims data the specificity was defined as:

$$Sp = \frac{n_{22}}{N_{2+}} = \frac{TN}{TN + FP}$$

Sensitivity and specificity are the most widely used indices for the assessment of performance of a diagnostic test. It is not widely appreciated that the propensity of a diagnostic test to generate particular proportions of positive and negative results means that non-zero levels of sensitivity and specificity are to be expected even if the classifications made by the diagnostic test are independent of the event. If a test is operating at 'chance' levels of performance, it is easy to demonstrate that the expected sensitivity is equal to the proportion of observations classified as positive by the test. Similarly, the expected specificity is the proportion of observations classified as negative. Kraemer [Kraemer HC and Bloch DA, 1994; Kraemer HC, 1992] has refined this notion, proposing two indices reflecting the 'quality' of sensitivity and specificity of the test. These indices adjust the raw sensitivity and specificity values for the values that would be expected to occur by chance alone. They may be seen as analogous to Cohen's kappa. Using the spreadsheet these quality indices were also estimated.

#### 4.5.4.3.4 Predictive Values

The conditional probabilities of observations that are positive and negative on the diagnostic test having or not having the event are referred to as the predictive values of positive (*PPV*) and negative (*NPV*) tests, respectively. These probabilities are also of use in assessing the performance of a test. They can be estimated as:

$$PPV = \frac{n_{11}}{N_{+1}} = \frac{TP}{TP + FP} \qquad NPV = \frac{n_{22}}{N_{+2}} = \frac{TN}{TN + FN}$$

#### 4.5.4.3.5 Likelihood Ratio

The likelihood or risk ratios (*LR*) of positive and negative tests and the overall odds ratio of the test were also calculated in our analysis. The likelihood ratios are measures of discrimination by a test result. These quantities do not depend on prevalence and provide the same information as sensitivity and specificity or the predictive values of the test but do so in a different form. A test result with a higher likelihood ratio raises the probability of detecting the event under study. The likelihood ratios can be defined as:

$$LR+ = \frac{Se}{1 - Sp} \qquad LR- = \frac{1 - Se}{Sp}$$

One can use the **LR** to represent rule-in or rule-out potentials. The rule-in and rule-out potentials as characterized by the **LR** is the “after-test” potential (post-test probability). That is, it is useful only *after* the test result is revealed. Using the **LR** of positive and negative tests, one can simply calculate the **Odds ratio** by dividing the likelihood of the positive test (**LR+**) by the likelihood of the negative test (**LR-**). The greater the odds ratio, the better predictive power the test possesses.

#### 4.5.4.3.6 Kappa

The kappa statistic ( $\kappa$ ) is another measure of overall agreement that is frequently used in the Health Services literature to summarize agreement between data sources [Kashner TM, 1998; Fowles JB, 1997; Romano PS and Mark DH, 1994; Jollis JG, 1993; Hannan EL, 1992; Horner RD, 1991]. The kappa statistic (weighted or unweighted) is appealing because it is a single index of agreement that considers chance. That is,  $\kappa$  statistics adjusts the agreement for rare conditions by using the potential agreement beyond chance as a baseline. It is defined as follows:

$$\kappa = \frac{\frac{n_{11} + n_{22}}{N} - \sum_{i=1}^2 \frac{N_{+i}N_{i+}}{N^2}}{1 - \sum_{i=1}^2 \frac{N_{+i}N_{i+}}{N^2}}$$

In the above formulation,  $\frac{n_{11} + n_{22}}{N}$  is the obtained agreement (*efficiency*) between the gold standard and the test under study, and is equal to the proportion correct coefficient. Also,  $\sum_{i=1}^2 \frac{N_{+i}N_{i+}}{N^2}$  is the agreement that would be expected by chance if the classifier was assigning the presence or the absence of an event at random with probabilities equal to the column proportions. Chance agreement assumes that the rows and column are independent. The definition of kappa is then *the agreement obtained beyond that which would be expected by chance compared to the maximum possible agreement that could be obtained*. Simply,

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} = \frac{P_o - P_e}{1 - P_e}$$

However, interpretation of  $\kappa$  is not always straightforward because this statistic is affected by prevalence [Berry CC, 1992; Cicchetti DV and Feinstein AR, 1990; Feinstein AR and Cicchetti DV, 1990]. For example, high levels of agreement between claims and hospital discharge summary data may emerge with low values of  $\kappa$  if the prevalence of the event of interest is low. In general, at extremes of prevalence kappa tends to decrease for a fixed sensitivity and specificity. However, to allow for comparability with other studies that measure agreement between claims and hospital discharge summary,  $\kappa$  values were reported. Kappa statistics, adjusted for bias and prevalence are known as *Scott's (pi) index* and *Bennett's agreement index*, respectively [Byrt T, 1993]. Lastly, 95% confidence intervals (CI) for each unadjusted

kappa were calculated using the usual method that is based on the asymptotic distribution theory of the maximum likelihood estimator [Kraemer HC and Bloch DA, 1994]. A wide 95% CI for kappa should serve as a cautionary note to the reader and is often, but not always, associated with disparate values of positive and negative agreement proportions.

Furthermore, alternative statistics not affected by the prevalence were also generated. For instance, Kullback-Leibler distance [Lee WC, 1999] can be used as an alternative measure for selecting diagnostic tests for ruling-out or ruling-in of an event.

#### **4.5.4.3.7 Kullback-Leibler Distance (KLD)**

The Kullback-Leibler distance (also referred to as ‘divergence’ or ‘information number’) is an abstract concept arising from statistics and information theory [Lee WC, 1999; Soofi ES, 1994]. It measures the distance or separation between two probability distributions, for example, diseased and disease-free distributions. This statistical measure is always non-negative and non-symmetric, with a large value indicating greater separation in the two distributions. It is zero if and only if the subjects in the two groups are identically distributed according to the distance test

[Soofi ES, 1994]. It is of interest to note that the Kullback-Leibler distances can be interpreted as the “before-test” potentials of ruling in and ruling out events.

For a binary diagnostic test, the Kullback-Leibler distance and the Rule-in/out probabilities can be defined by the following equations, where the proportions of the events and the non-events are assumed known and are denoted as  $f$  and  $g$ , respectively.

$$D(f \parallel g) = (1 - Se) \cdot \log \frac{1 - Se}{Sp} + Se \cdot \log \frac{Se}{1 - Sp}$$

$$= (1 - Se) \cdot \log LR_- + Se \cdot \log LR_+$$

$$P_{Rule-in} = \exp[D(f \parallel g)]$$

$$D(g \parallel f) = Sp \cdot \log \frac{Sp}{1 - Se} + (1 - Sp) \cdot \log \frac{1 - Sp}{Se}$$

$$= Sp \cdot \log(1 / LR_-) + (1 - Sp) \cdot \log(1 / LR_+)$$

$$P_{Rule-out} = \exp[D(g \parallel f)]$$

#### 4.5.4.3.8 Levels of Agreement

For the purpose of this thesis, proportion that constitutes a satisfactory level of agreement should be defined. In clinical studies, percentage agreement greater than 50% has been noted as ‘moderate’, agreement greater than 75% as ‘substantial’,

agreement greater than 90% as 'almost perfect', and 95% and more as 'excellent or perfect'. This would provide a preliminary approach to comparison [Szklo.M and Nieto FJ, 2000].

This is in accordance to some experts who "have attached the following qualitative terms to Kappa: 0.0-0.2 = 'slight', 0.2-0.4 = 'moderate', 0.6-0.8 = 'substantial', and 0.8-1.0 = 'almost perfect'" in the context of clinical epidemiology [Sackett DL, 1991]. In broad terms a kappa below 0.2 indicates poor agreement and a kappa above 0.8 indicates very good agreement beyond chance [Landis JR and Koch GG, 1977a; Landis JR and Koch GG, 1977b]. These levels of agreement were also considered when analyzing the quality of data in the RAMQ medical claims database.

### 4.5.5 Application

Validity of an algorithm developed for a specific clinical condition should be assessed before being applied to other conditions. In so doing, one could examine whether the model or the algorithm is a good description of the validation data set, in terms of its behavior and of the application proposed (that is, the *generalizability* of the algorithm).

Assessment of generalizability can be defined as the evaluation of consistency of the results of an algorithm (i.e., the algorithm form together with the parameter estimates) obtained from a learning or index data set with a validation data set not used for development of the algorithm and its estimation. This process directly depends on the objective of the analysis. Not all analytical algorithms may need to be generalized. However, *their predictive performance* should be examined to verify the stability of the parameter estimates in various conditions (such as study designs, clinical indications, etc.).

Originally, the algorithm was developed for the assessment of the data quality in the RAMQ medical services database and analysis of data pertaining to the hospital morbidity indices of COPD. However, the algorithm must be examined whether it could be applied (or extrapolated) to other diseases.



Basically, there are two types of validation study that can be used to generalize the result or the application of a project: *external* and *internal*. External validation is the application of the developed algorithm to a new data set (validation data set); whereas, internal validation refers to the use of data-splitting and resampling techniques (e.g., cross-validation and bootstrapping). Since internal validation techniques are beyond the scope of this thesis, they will not be herein discussed. External validation, indeed, provides the most stringent method for testing the developed algorithm.

In our study, validation was performed by applying the algorithm to a new data sample received from the RAMQ. Although diagnostic criteria for patients' selection differed from the previous study, the ability of the developed algorithm in estimating episodes of hospitalizations in the RAMQ medical services database—regardless of diagnosis at admission—was first examined. The robustness of the algorithm was then assessed for two new clinical conditions: acute myocardial infarction (AMI) and upper gastrointestinal bleeding or perforation (UGIB).

#### **4.5.5.1 Study Sample**

Medical, dispensed prescription and demographic records were obtained from the RAMQ database for all patients (707,916 individuals) aged 65 years or older who

had received a dispensed prescription for Aspirin, Acetaminophen, NSAIDs, or COX-2 inhibitors between April 1, 1999 and December 31, 2002.

Hospital discharge summary of the subjects in the RAMQ database, regarding acute myocardial infarction (AMI) and upper gastrointestinal bleeding (UGIB) were provided by the MED-ECHO for the same study period; there was an additional opportunity to assess the accuracy and validity of ICD-9 diagnostic codes indicating episodes of hospitalizations for these two clinical conditions.

RAMQ and MED-ECHO data provided for this study were devoid of any information that could lead to the identification of individuals and were limited to only variables required for the study analyses.

#### **4.5.5.2 Analyses**

##### **4.5.5.2.1 Primary**

Using the RAMQ medical services database, inhospital claims were identified using the institutional codes. Inhospital morbidity indices (LOS and number of hospitalizations) were estimated by applying the developed algorithm to the RAMQ database, regardless of the diagnosis at admission. For the same period, all relevant hospitalizations for our target population were identified using MED-ECHO database.

Robustness of the algorithm to estimate episodes of hospitalizations was then assessed using descriptive statistics, by which parameter estimates obtained from the RAMQ were compared to that of observed in MED-ECHO database as the gold standard.

#### **4.5.5.2.2 Secondary**

The RAMQ Claims data from inpatient records were retrieved if the claims were one of the ICD-9 codes for AMI (410.X) or UGIB (531X-534X, 578X). Subsequently, in two separate analyses, validity of ICD-9 diagnostic codes in the RAMQ claims database to identify episodes of hospitalizations for AMI and UGIB were assessed.

Inhospital claims of individuals who suffered an acute myocardial infarction were identified using disease-specific inclusion and exclusion criteria. This procedure was simple and straightforward, since only claims with the diagnosis '410' had to be selected. Post admission comorbidities and complications were not considered. Those AMI cases with lengths of stay less than 3 ( $LOS < 3$ ) days who were discharged alive were excluded since "rule-out" myocardial infarction has been a source of misclassification in prior studies [Petersen LA, 1999; Schiff GD and Yaacoub AS, 1989; Iezzoni LI, 1988]. Occasional acute bed shortages may have dictated early discharge of some valid AMI patients. Under these conditions, physicians are under

intense pressure to release patients. Clarification of this issue requires chart reviews of these cases or a formal model incorporating hospital capacity utilization, which could be used to determine whether non-financial aggregate demand factors influence discharge decisions. However, these criteria are strongly recommended for use in comparisons of AMI treatment and outcomes [McClellan M and Kessler D, 1999; Tunstall-Pedoe H, 1988]. It should be noticed that episodes of care for AMI are defined beginning with the day of the index admission (i.e., the first hospitalization record with at least 2 physician claims filed with the diagnosis code '410').

However, to identify episodes of hospitalizations for UGIB, patients were selected according to different inclusion and exclusion criteria using combinations of different diagnoses and/or procedural codes for upper gastrointestinal endoscopy, which resulted in the development of 4 models (**M1-M4**). The following is the description of these models:

- ⇒ **M1**: First, potential hospitalizations for UGIB were identified if they had a claim for upper gastrointestinal endoscopy within 24 hours of their admission to the hospital (admission date  $\pm$  1 day). Because it is most likely that a patient undergoes endoscopy if there is an indication of UGIB. All endoscopy claims in the emergency rooms, in addition to those filed with in-hospital codes, were considered for this analysis.

- ⇒ **M2:** In other analysis, potential cases for UGIB were identified if they had a physician claim with the following ICD-9 diagnostic codes: gastric ulcer (531.X), duodenal ulcer (532.X), peptic ulcer (533.X), gastrojejunal ulcer (534.X), and gastrointestinal hemorrhage (578.X).
- ⇒ **M3:** From the previous model (M2), we later excluded hospitalizations that contained a code for gastrointestinal hemorrhage (ICD-9 578.X) and had one or more codes for a neoplasm of the gastrointestinal tract (ICD-9 151.X, 153.X, 154.X, 197.5, 211.3), esophageal varices (ICD-9 456.0, 456.2), gastroesophageal laceration-hemorrhage syndrome (ICD-9 530.7), acute vascular insufficiency of intestine (ICD-9 557.0), diverticulosis or diverticulitis of colon with hemorrhage (ICD-9 562.1), angiodysplasia of intestine with hemorrhage (ICD-9 569.8), or other specified disorders of rectum and anus (ICD-9 569.4).
- ⇒ **M4:** Finally, hospitalizations that met the criteria for model M3 and had a claim for upper gastrointestinal endoscopy within the first 24 hours of their admissions were selected for the analysis.

All in-hospital claims of individuals who suffered from an acute myocardial infarction or an upper gastrointestinal bleeding or perforation between April 1<sup>st</sup>, 1999 and March 31<sup>st</sup>, 2002 were separately identified, using the ICD-9 diagnostic codes

and/or procedural codes on the RAMQ database—**inhospital claims file**. Hospital episodes were also identified by applying the developed algorithm (with an optimal gap) to the RAMQ medical claims database for the same study period—**hospitalizations file**.

By linking the *inhospital claim and hospitalizations files*, hospital episodes containing any of the ICD-9 code maps described above were identified for all patients over 65 years of age during the specified period for AMI and UGIB, independently. In other words, the algorithm was designed in order that any patient with an inhospital physician service claims bearing a diagnosis of AMI or UGIB at any time during hospitalizations (within admission-discharge events) was denoted as having an AMI- or a UGIB-related hospitalization, respectively. This procedure ensured the identification of all potential hospitalizations for the two clinical conditions of interest in the RAMQ medical claims data, which would result in a higher sensitivity of the algorithm identifying an ICD-9 code (i.e., to increase the proportion of common hospitalizations that were identified by the two databases).

Using MED-ECHO database and based on the same case definitions (ICD-9 diagnostic codes), all discharged hospitalizations with the diagnostic codes of AMI and UGIB were identified for the entire study period. For the purpose of homogeneity, hospitalization records with LOS < 3 days where patients were discharged alive were also excluded. Records from the RAMQ medical claims and MED-ECHO discharge summary were matched by unique patient identification numbers and hospitalization

dates. Only the patient's first hospitalization was included in the agreement analysis since the unit of analysis was considered to be "patient" rather than the "number of hospitalizations". Otherwise, there would be a discrepancy in the unit of analysis of true negative (the TN cell) as compared to the other three cells (i.e., TP, FP, and FN cells) in the 2x2 contingency table. A true negative would indicate that the patient did not have any hospitalization record in either database (i.e., RAMQ and MED-ECHO), whereas in the case of the true positive, false positive, and false negative there might be more than one hospitalization per patient. To be consistent in all the cells, the patient's first hospitalization was included in the agreement analysis so that it would indicate the number of patients rather than the number of hospitalizations.

The central tendency measures of the estimated parameters using the RAMQ were finally compared to those observed in MED-ECHO database. Based on the clinical condition and the ICD-9 map described earlier, two-by-two tables were constructed to assess the agreement between the data sources. If a hospitalization was present in the two databases it was considered to be a "true positive", whereas it was considered a "true negative", if it was absent from the two databases. The "false negative" and the "false positive" cells were represented by identifying a missing hospitalization from either the RAMQ or the MED-ECHO database, respectively. Validity was assessed by calculating sensitivity, specificity, efficiency, and positive and negative predictive values; and more specific measures, such as Kullback-Leibler index, when compared to the gold standard database, MED-ECHO. Kappa statistics

were generated for each condition to measure reliability and data agreement while controlling for chance agreement. Confidence intervals and test statistics for proportions were calculated by the normal approximation.

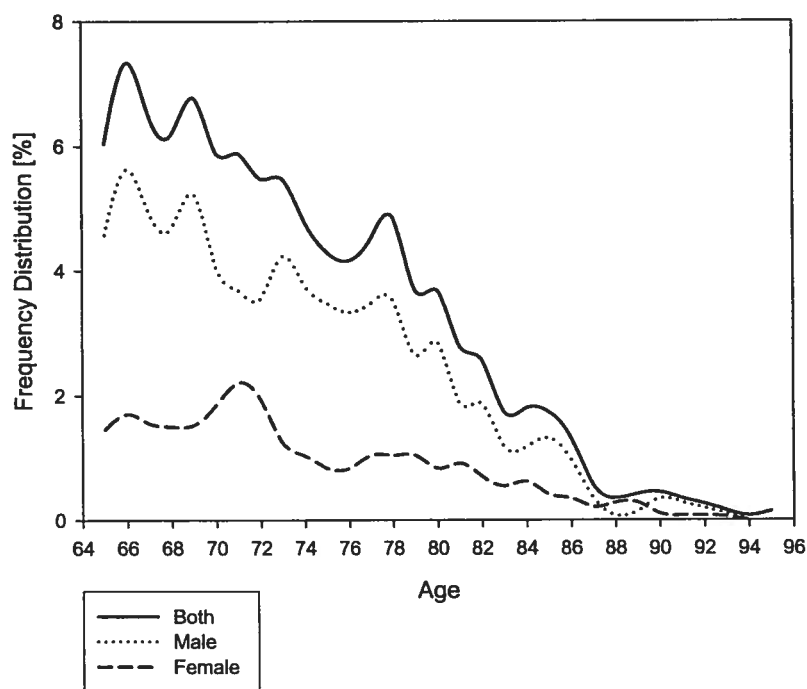


## **5.1 Developed Algorithm (COPD)**

Using RAMQ administrative database, 3,768 COPD-diagnosed patients had previously been identified (either at a hospital, emergency department, or medical clinics), who had chronic use of bronchodilators between January 1<sup>st</sup>, 1990 and June 30<sup>th</sup>, 1996 [Blais L, 2004]. Within the initial cohort and the use of RAMQ medical services claim database, only patients who had at least one physician claim for medical services, including inpatient and outpatient claims, were identified (N=1,441). Among these patients, only individuals (N=1,233) who were admitted to Quebec hospitals at least once during the study period of January 1, 1990, to December 31, 1996 were selected for the final analyses. According to the RAMQ medical service claims database, a total of 146,170 inpatient services were provided to these patients.

In the initial cohort, 75% of the COPD-diagnosed patients were males, and the mean age at cohort entry was  $74 \pm 6$  years (mean  $\pm$  SD). Distributions of age and sex were similar in the sub-cohort of individuals with in-hospital service claims. Figure 5.1 illustrates the age distribution of COPD-diagnosed patients in males and females

separately and combined between 1990 and 1996. As shown in this Figure, the prevalence of COPD was found to be age-dependent.



**Figure 5.1** – Age distribution in the COPD cohort by gender.

### 5.1.1 Central Tendency

Episodes of hospitalizations were estimated for this period by applying the developed algorithm using different gaps (period of service interruption within a unique hospitalization was allowed to vary between 1 and 15 days). Inhospital length

of stay (LOS) and number of hospitalizations derived from RAMQ were then compared to that of observed in MED-ECHO. Table 5.1 and Table 5.2 show the central tendency (i.e., mean, median, etc.) of the inhospital LOS and number of hospitalizations obtained from the RAMQ medical claims using the algorithm with different gaps compared to that of observed in the gold standard.

The total number of hospitalizations for the target population as identified by the gold standard (MED-ECHO database) was found to be 4,615 hospitalizations (average of  $3.7 \pm 3.1$  [ $\pm$  SD]); and an average length of stay of  $13.5 \pm 26.6$  [ $\pm$  SD] days. Nevertheless, after applying the algorithms (with different gaps) on the RAMQ medical claims database, the total number of hospitalizations varied between 9,767 when the gap of 1 day was considered and 4,089 hospitalizations for the gap of 15 days. It was noticed that increasing the gap from 1 day to 15 days resulted in decreasing the average number of hospitalizations (from 11.9 to 3.2 times), whereas the average inhospital LOS increased (ranged between 4.6-14.6 days); however, the extent to which these changes occurred were not evident after the gap of 7 days.

The appropriateness of various standard measures of central tendency for analyzing inhospital LOS and number of hospitalizations was also explored. The averages of the inhospital length of stay and the number of hospitalizations of all patients were significantly higher than the medians, reflecting the rightward skewness. Patients with extended stays heavily influenced mean length of stay.

**Table 5.1:** Estimated LOS in the RAMQ Medical Claims Database.

	N	Mean $\pm$ SD	Median	95% CI	CV%
<b>Length of Stay</b>					
<b>Gold</b>	4615	13.5 $\pm$ 26.6	7.0	13.0 — 14.6	193%
<b>Gap (days)</b>					
1	9767	4.6 $\pm$ 6.3	2.0	4.4 — 4.7	139%
2	7218	6.5 $\pm$ 8.4	4.0	6.3 — 6.7	129%
3	5861	8.5 $\pm$ 11.2	5.0	8.2 — 8.8	131%
4	5312	9.7 $\pm$ 13.1	6.0	9.3 — 10.0	135%
5	5039	10.4 $\pm$ 14.7	6.0	10.0 — 10.8	141%
6	4827	11.1 $\pm$ 15.8	7.0	10.7 — 11.5	142%
7	4562	12.2 $\pm$ 17.7	7.0	11.6 — 12.6	147%
8	4461	12.5 $\pm$ 18.9	7.0	12.0 — 13.1	151%
9	4396	12.8 $\pm$ 19.8	7.0	12.2 — 13.4	155%
10	4323	13.2 $\pm$ 21.1	8.0	12.6 — 13.8	160%
11	4772	13.5 $\pm$ 21.4	8.0	12.8 — 14.1	159%
12	4217	13.8 $\pm$ 22.6	8.0	13.1 — 14.5	164%
13	4176	14.0 $\pm$ 23.3	8.0	13.3 — 14.8	166%
14	4136	14.3 $\pm$ 23.8	8.0	13.6 — 15.0	166%
15	4089	14.6 $\pm$ 24.2	8.0	13.9 — 15.4	166%

\* CI: Confidence Interval.

\*\* CV: Coefficient of Variation.

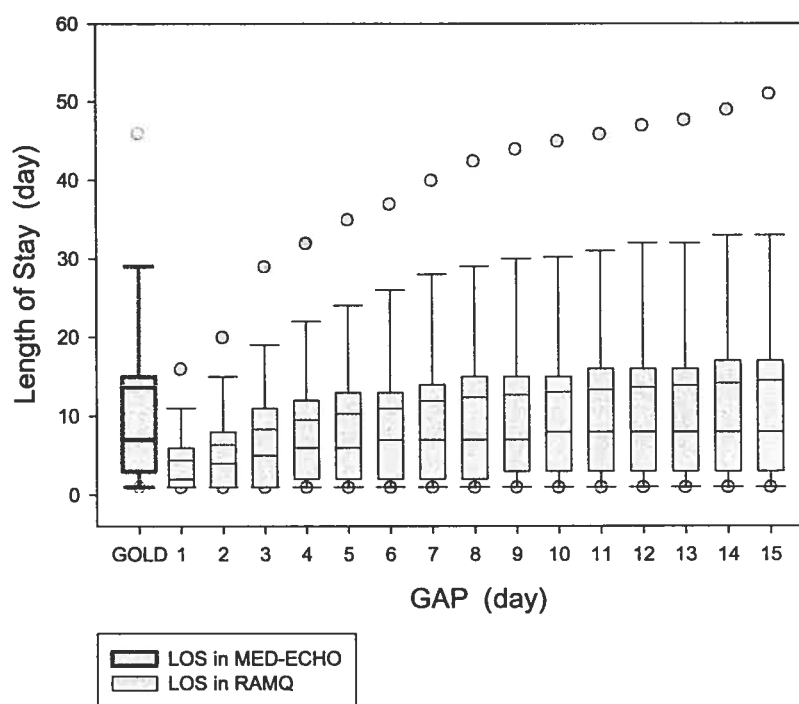
**Table 5.2:** Estimated N° of Hospitalizations in the RAMQ Medical Claims Database.

	<b>N</b>	<b>Mean ± SD</b>	<b>Median</b>	<b>95% CI</b>	<b>CV%</b>
<b>N° of Hospitalizations</b>					
<b>Gold</b>	4615	3.7 ± 3.1	3.0	3.6 — 3.8	83%
<b>Gap (days)</b>					
<b>1</b>	9767	11.9 ± 15.3	6.0	11.6 — 12.2	128%
<b>2</b>	7218	7.7 ± 9.4	4.0	7.5 — 7.9	122%
<b>3</b>	5861	5.5 ± 5.9	4.0	5.3 — 5.6	107%
<b>4</b>	5312	4.7 ± 4.6	3.0	4.5 — 4.8	99%
<b>5</b>	5039	4.3 ± 4.2	3.0	4.2 — 4.7	96%
<b>6</b>	4827	4.1 ± 3.8	3.0	4.0 — 4.0	93%
<b>7</b>	4562	3.6 ± 3.0	3.0	3.6 — 3.7	81%
<b>8</b>	4461	3.6 ± 2.9	3.0	3.5 — 3.6	81%
<b>9</b>	4396	3.5 ± 2.8	3.0	3.4 — 3.6	80%
<b>10</b>	4323	3.4 ± 2.7	3.0	3.4 — 3.5	79%
<b>11</b>	4772	3.4 ± 2.7	3.0	3.3 — 3.5	79%
<b>12</b>	4217	3.3 ± 2.7	3.0	3.3 — 3.4	79%
<b>13</b>	4176	3.3 ± 2.6	2.0	3.2 — 3.4	79%
<b>14</b>	4136	3.3 ± 2.6	2.0	3.2 — 3.4	79%
<b>15</b>	4089	3.2 ± 2.6	2.0	3.2 — 3.3	79%

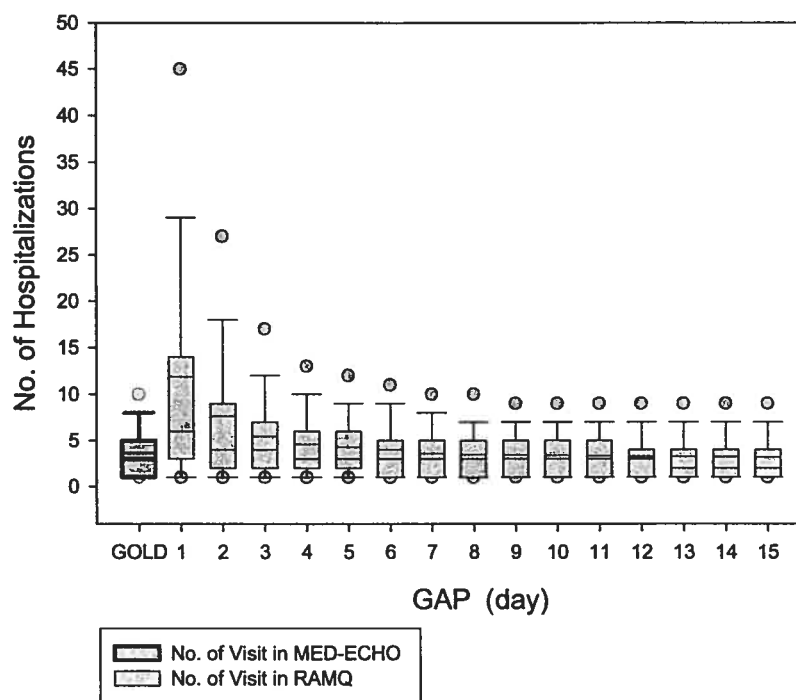
\* CI: Confidence Interval.

\*\* CV: Coefficient of Variation.

Figure 5.2 and Figure 5.3 demonstrate the parameter estimates in the RAMQ data compared to MED-ECHO as the gold standard. In the box plots, ranges or distribution characteristics of the selected variables (i.e., LOS and N° of hospitalizations) are plotted separately for the observed values in the gold standard and the estimates derived from the RAMQ data using different gaps. The central tendency (i.e., mean and median), midrange (hinges of the boxes), and the data range between 5% and 95% quartiles (whiskers) are computed for each group of cases and the selected values are presented. Because of the long-tailed distribution (heavy skewness), data points beyond 5% to 95% range are not shown.



**Figure 5.2** – Estimated length of stay (LOS) in RAMQ using different gaps.



**Figure 5.3** – Estimated N° of hospitalizations in RAMQ using different gaps.

Visual analysis of the frequency distributions (Figure 5.2 and Figure 5.3) also indicated that the medians and modes reflected the central tendency better than the mean.

The tables of results, as well as the graphs, of the parameter distributions demonstrated that the distribution of the in-hospital LOS and the number of hospitalizations in the target population were skewed to the right. Visual analyses of graphs were useful in determining trends quite rapidly. The simple calculation of the coefficient of variation (i.e., a descriptive statistic defined as the ratio of the standard

deviation of a measure to the mean of the measure) also helped identify some of the circumstances where visual inspection of the distribution graph was needed.

### 5.1.2 Predictive Performance

Visual analyses of the graphs also revealed that the episodes of hospitalizations (inhospital LOS and N° of hospitalizations) can be well defined using the analytical algorithm with a gap approximately equal to 7 days. Therefore, predictive performance analyses were carried out in order to discover the algorithm (with an optimal value for the gap) that best estimates and well describes the inhospital LOS and Number of hospitalizations in our target population.

Table 5.3 shows the results of the absolute predictive performance analyses on the lengths of stay of all hospitalizations for selected algorithms, although the analyses were done for all of them. Only those that were shown to provide close estimates in visual analyses were included in this table. Bias and accuracy of the parameter estimates were computed as mean error (**ME**) and mean absolute error (**MAE**), respectively. The precision was calculated as root mean squared error (**RMSE**). Moreover, cumulative error of the stream of predictions (percent deviation) for the episodes of hospitalizations was estimated as percent prediction error (**%PE**) and percent absolute prediction error (**%APE**), weighted by predicted values; distribution of percent prediction error is graphically shown in Figure 5.4. Normally distributed,



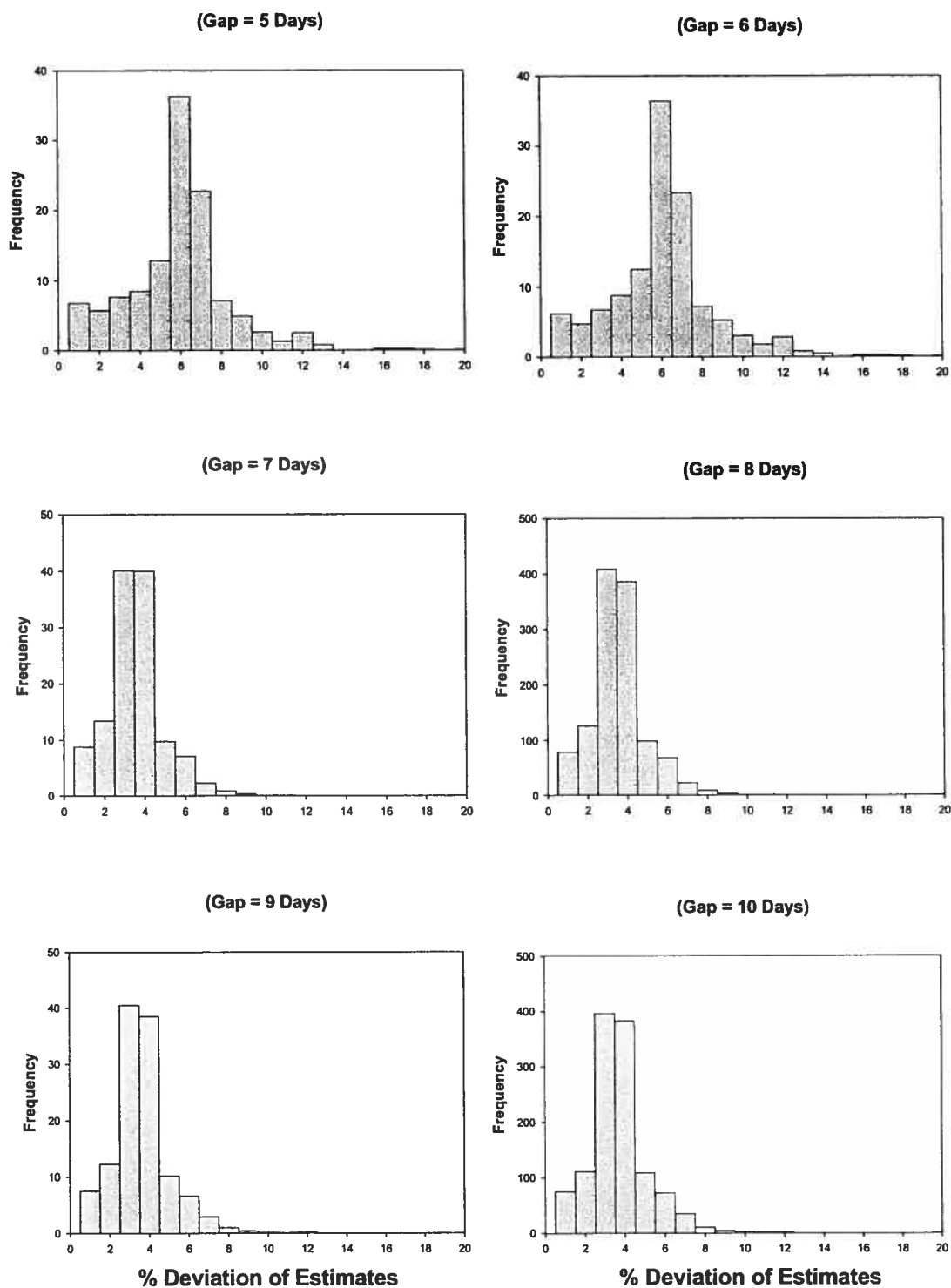
the graph representing the prediction error using 7 days of gap clearly shows a balance between over- and under-estimated values for the average lengths of stay.

**Table 5.3:** Absolute predictive performance of selected algorithms (gaps).

<b>GAP</b> <b>(N° of Days)</b>	<b>ME</b> (95%CI)	<b>MAE</b> (95%CI)	<b>%PE</b> (95%CI)	<b>%APE</b> (95%CI)	<b>RMSE</b> (95%CI)
<b>4</b>	3.8 (2.8–4.8)	5.4 (4.5–6.4)	-7.9 (-10.3–-5.4)	31.6 (29.8–33.4)	19.0
<b>5</b>	3.1 (2.2–4.1)	5.1 (4.2–6.0)	-4.1 (-6.6–-1.6)	30.8 (28.9–32.6)	17.9
<b>6</b>	2.6 (1.1–3.5)	4.9 (4.1–5.8)	-0.9 (-3.5–1.7)	30.8 (28.9–32.7)	17.3
<b>7</b>	1.8 (1.0–2.6)	4.4 (3.7–5.2)	2.8 (0.1–5.6)	26.3 (24.2–28.3)	14.9
<b>8</b>	1.3 (0.6–2.1)	4.6 (3.8–5.3)	5.0 (2.2–7.8)	26.6 (24.6–28.7)	15.8
<b>9</b>	1.0 (0.2–1.8)	4.5 (3.8–5.3)	7.0 (4.1–9.9)	33.0 (30.7–35.3)	15.2
<b>N</b>	1441	1441	1441	1441	1441

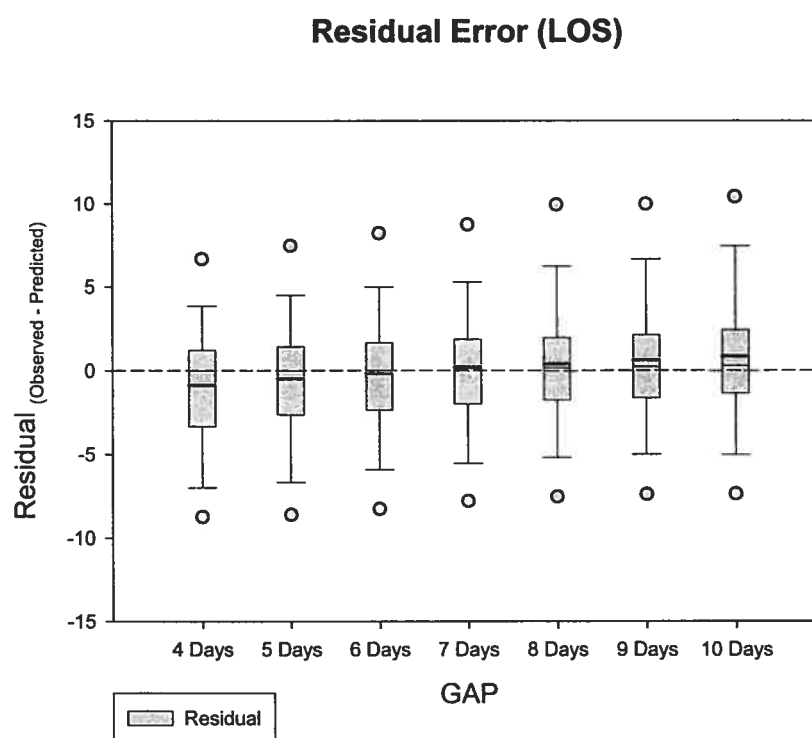
⇒ **ME** (mean error); **MAE** (mean absolute error); **RMSE** (root mean squared error);  
**PE%** (percent prediction error); **APE%** (percent absolute prediction error).

⇒ **95%CI** (95% confidence interval)

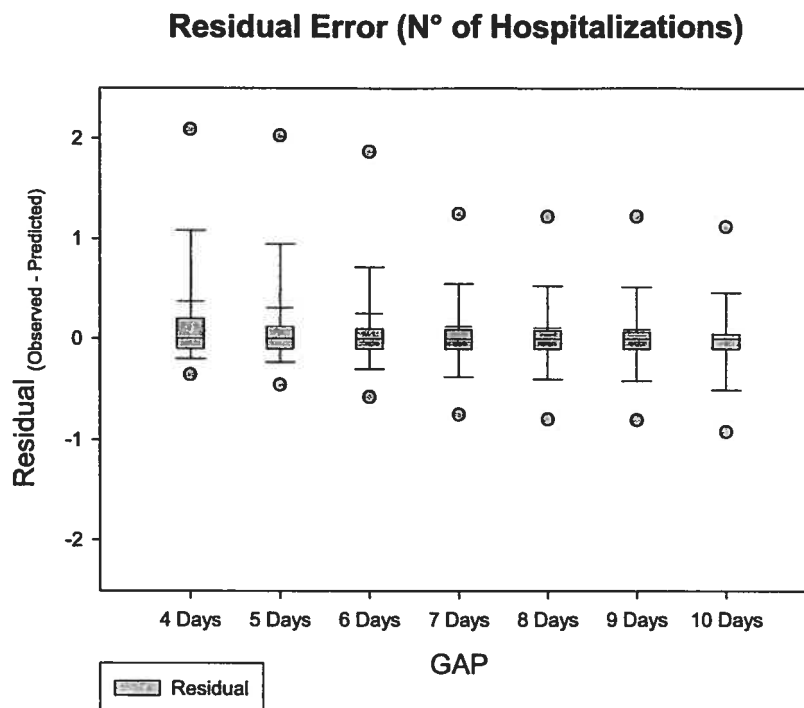


**Figure 5.4** – Distribution of predictive performance (% PE) for selected algorithms.

Figure 5.5 and Figure 5.6 demonstrate the residual error of estimates for length of stay and number of hospitalizations, respectively. Boxplot presentations were used to better compare the results of different algorithms in terms of mean and median predictive values. Predictive performance measures are presented as the mean values with 95% confidence interval (95%CI). As shown in these figures, the residual errors for the lengths of stay and number of hospitalizations were smaller for the algorithm with the 7 days of gap, where the error dispersion followed Gaussian distributions (equal mean and median for residual estimates) with the tightest confidence limits, even when compared to the closest algorithms (i.e., the gap of 5 and 6 days).



**Figure 5.5** – Residual error estimates of selected algorithms (LOS).



**Figure 5.6** – Residual error estimates of selected algorithms (N° of Hospitalizations).

Based on the results obtained from the predictive performance analyses, the least biased estimates were achieved using the gap of 9 days. In fact, the bias in estimation significantly improved when the gap increased from 4 to 9 days. However, the best accuracy in predictions (MAE) was found when the gap was set to 7 days in our algorithm with a balanced prediction for over- and under-estimated values. Altogether, the analyses of bias and accuracy tended toward the algorithm with the gap of 7 days. Furthermore, the estimates of percent prediction error (%PE) and percent absolute prediction error (%APE) pointed to 7 days as the optimal gap; since

they provided not only the smallest values, but also they were normally distributed among the predictions (Figure 5.4). Finally, the RMSE (as the indication for precision) was computed for different algorithms; it appeared to be the lowest value for the gap of 7 days. The smallest value of RMSE indicated that the most precise algorithm to estimate episodes of hospitalizations was the one with the gap of 7 days.

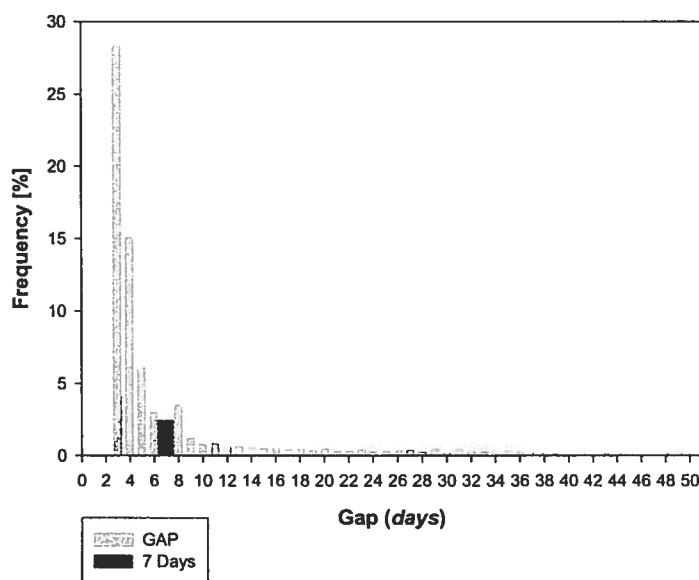
Overall, the tables of results and graphical presentations of the estimated parameters, as well as the predictive performance analysis revealed that the developed algorithm with the gap of 7 days well defined inhospital length of stay and number of hospitalizations in our target population. However, further analysis was required to justify the use of this algorithm (gap=7 days) for identification of episodes of hospitalizations, since the gap of 7 days seemed to be unreasonable. Thus, the service interruption (the gap) in the RAMQ medical claims database and the interval between hospitalizations in the MED-ECHO database were obtained in our target population for the whole study period.

Average length of gap in the RAMQ file was 87.9 days with the median and mode equal to 7 and 4 days, respectively; whereas average interval between hospitalizations in MED-ECHO appeared to be 225.3 days with the median and mode of 110 and 8, respectively.

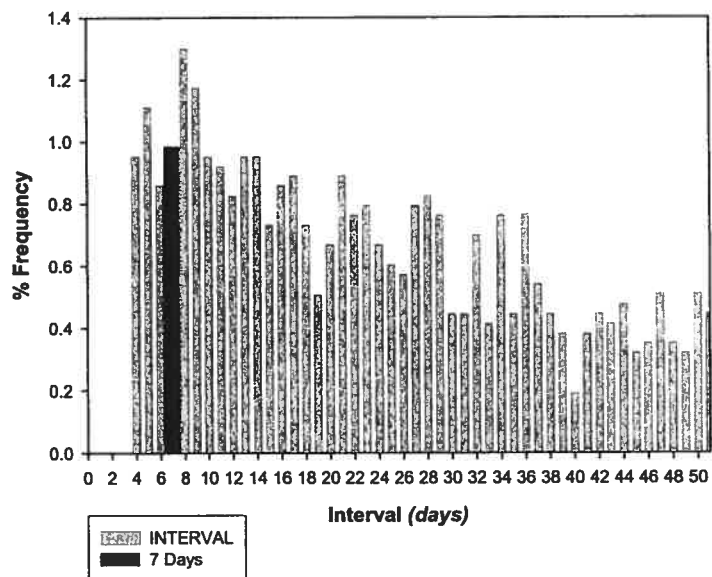
Since the distributions of these values were greatly skewed to the right, one should not rely on the average values for comparison. The median gap in the RAMQ

roughly represented the average time by which half of the patients had not received any in-hospital services. This gap could be either the time between two hospitalizations or the time within a unique hospitalization. Comparing the median gap in the RAMQ to the central tendency of the hospitalization intervals in MED-ECHO indicated that even considering a long service interruption (a gap of 7 days) would not result in the overestimation of the LOS and/or underestimation of the number of hospitalizations in our target population, because considering the gap of 7 days would not even overlap with the most frequent interval between hospitalizations (mode=8 days).

Figure 5.7 demonstrates the distribution of the gap in the RAMQ medical claims database, whereas Figure 5.8 represents the distribution of hospitalization intervals in MED-ECHO.



**Figure 5.7** – Frequency distribution of gaps in the RAMQ medical claims database.



**Figure 5.8** – Frequency distribution of hospitalization intervals in MED-ECHO.

Consequently, based on the results obtained from all analyses, the analytical algorithm with the gap of 7 days was selected for more exploration and validation of data quality in the RAMQ medical services database (analysis of agreement). Episodes of hospitalizations were best defined using this algorithm.

Figure 5.9 and Figure 5.10, respectively, demonstrate histogram distribution and probability distribution function (pdf) of LOS and number of hospitalizations in the RAMQ medical claims database compared to that of obtained from MED-ECHO.

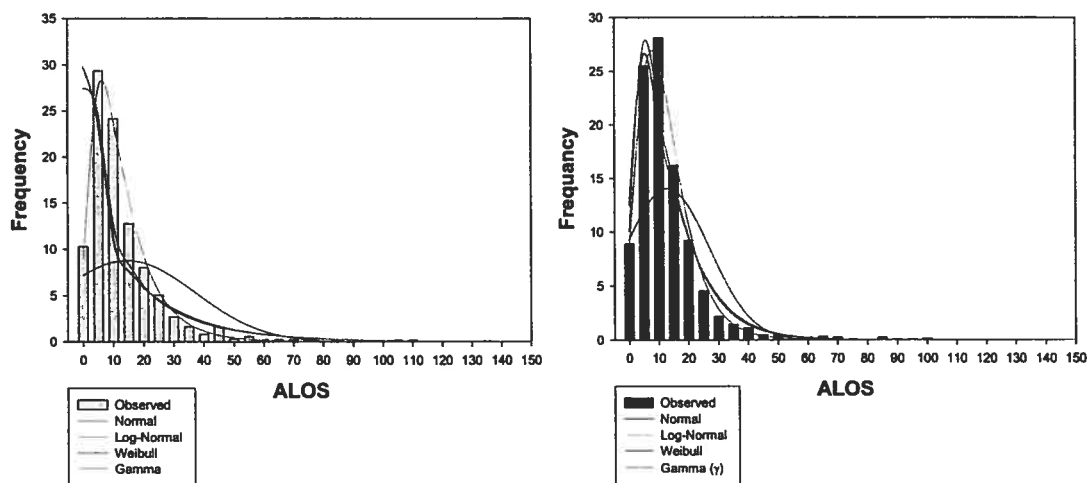


Figure 5.9 – Histogram of average LOS and its *pdf*.

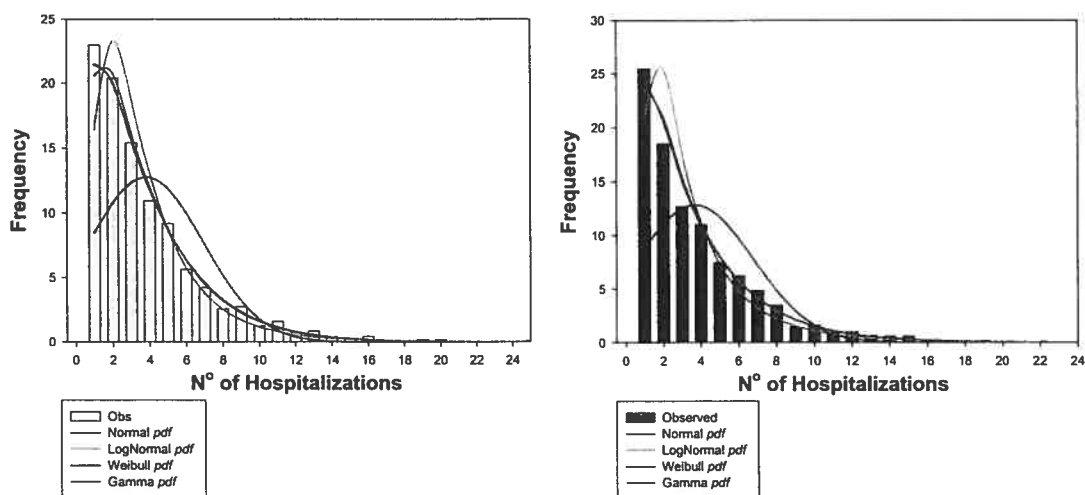
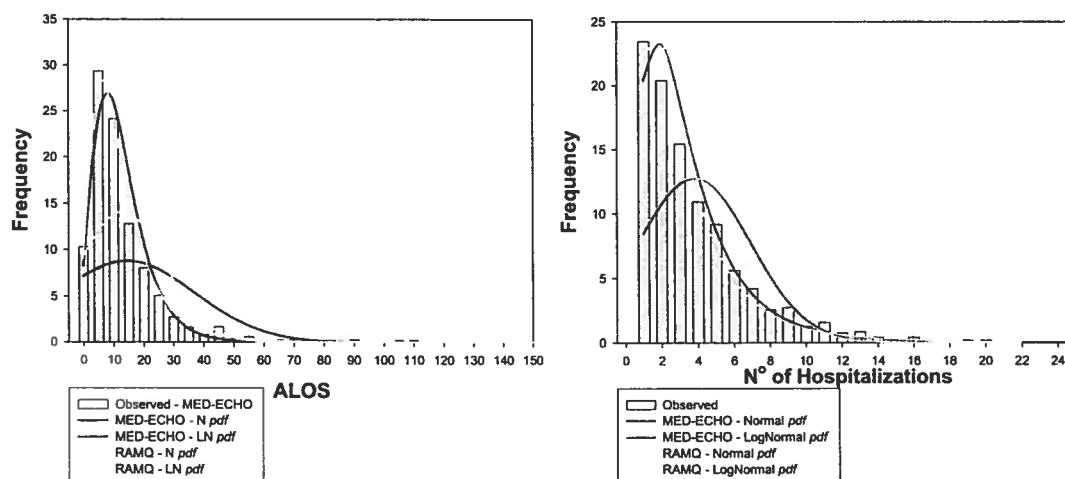


Figure 5.10 – Histogram of N° of hospitalizations and its *pdf*.



As shown in Figure 5.9 and Figure 5.10, distribution of LOS and number of hospitalizations in our target population can not be approximated using a Gaussian (normal) distribution. Both LOS and number of hospitalizations were well described using lognormal distribution (the values of median and geometric mean were almost equal).

Figure 5.11 shows the comparison of parameter estimates (LOS and number of hospitalizations) obtained from the RAMQ medical claims database to that of observed in MED-ECHO.



**Figure 5.11** – Comparison of the estimated LOS and N° of hospitalizations in the RAMQ to those observed in the gold standard (MED-ECHO). As shown in these graphs, LOS and number of hospitalizations both follow a lognormal distribution.

Table 5.4 shows the summary results obtained from the algorithm with the 7 days of gap in the RAMQ medical claims database, compared to that of observed in MED-ECHO. For the best algorithm (gap=7 days), the mean estimated LOS was 12.2 days with the median and geometric mean of 7.0 and 6.2 days in RAMQ, respectively; whereas the average LOS observed in the MED-ECHO file was 13.5 days with median and geometric mean of 7 and 6.7 days ( $p < 0.001$ ). The average readmission was estimated to be 3.7 versus 3.8 ( $p = 0.36$ ) for the entire study cohort in the RAMQ and MED-ECHO databases, respectively.

**Table 5.4** – Estimated population parameters in a cohort of COPD patients using RAMQ database, compared to the Gold Standard.

	RAMQ (Gap = 7 days) n = 1,233	MED-ECHO (Gold Standard) n = 1,206	P value *
<b>Length of Stay</b>			
Mean	12.2	13.5	< 0.001
Range	1 – 241	1 – 367	
Geometric Mean	6.2	6.7	< 0.001
Quantile:			
25% Q1	5.0	5.0	
50% Median	7.0	7.0	0.9901
75% Q3	17.0	15.0	
<b>No. of Hospitalizations</b>			
Mean	3.7	3.8	0.3618
Range	1 – 22	1 – 25	
Geometric Mean	3.0	3.0	< 0.001
Quantile:			
25% Q1	1.0	2.0	
50% Median	3.0	3.0	0.9506
75% Q3	5.0	5.0	
* : P value is estimated based on two-sided t-test, however, it is estimated based on median two-sample test for 50% Median estimates; n: number of patients.			

The geometric means were lower than the arithmetic means for both parameters; they tended to track the median values. The geometric mean was computed as the mean of the natural logarithm transformation. Log transformation did decrease the skewness of the frequency distributions and affects extremely long tails consisting of only a few patients per length of stay and number of hospitalizations.

### **5.1.3 Analysis of Agreement**

Among the target population (3,768 COPD-diagnosed patient), 1,206 individuals had been hospitalized at least once during the study period. The total number of hospitalizations for this cohort as identified by MED-ECHO database was found to be 4,615 hospitalizations with the cumulative length of stay of 63,610 days. Applying the developed algorithm on the RAMQ medical claims database, 1,233 individuals were identified to have at least one hospitalization for the same study period with the total number of hospitalizations and cumulative duration of 4,562 and 52,604 days, respectively. Among hospitalizations estimated from the RAMQ medical claims database, 3,727 hospitalizations were matched to that of MED-ECHO (approximately 81%) using a deterministic automated matching algorithm; whereas, 835 hospitalizations were not matched. Within group of hospitalizations that matched, the length of stay was found to be accurately estimated in 1,600 (42.9%)

hospitalizations, underestimated in 1,546 (41.5%), or overestimated in 581 (15.6%) hospitalization cases.

Compared to MED-ECHO, the sensitivity and specificity of the best algorithm identifying the first in-hospital stay (1,174 cases) were 92.5% (90.8-93.9%) [95%CI] and 95.4% (94.5-96.2%), respectively. The efficiency of the RAMQ medical services database for estimating episodes of hospitalizations was estimated to be 94.5% (93.7-95.2%) when compared to the MED-ECHO database as the gold standard. Moreover, the predictive value for the positive test was found to be 90.4% (88.7-92.0%), while it was 96.4% (95.6-97.1%) for the negative test.

In this study, the reliability (kappa coefficient) of the RAMQ data identifying COPD hospitalization episodes was found to be  $\kappa=0.87$  (95%CI: 0.86-0.89). This level of reliability indicated almost perfect agreement between these databases for estimating LOS and number of hospitalizations in the COPD cohort. Kappa statistics further improved (from 0.87 to 0.89) when it was adjusted for prevalence and bias; however, these adjustments did not change the magnitude of the reliability formerly achieved. For more details on diagnostic and agreement statistics please refer to Appendix 2, Table 2A.

## 5.2 Application

Validity and reliability of the developed algorithm in estimating episodes of hospitalizations for other clinical conditions and with different diagnoses at admission were then examined. Furthermore, the accuracy of the RAMQ medical claims data identifying hospitalization records for two new clinical conditions (i.e., acute myocardial infarction (AMI) and upper gastrointestinal bleeding (UGIB)) based on the diagnostic and/or procedural codes, was separately assessed. These processes were once more implemented through comparison analyses against MED-ECHO database (the gold standard) using an automated deterministic linkage method.

A retrospective, population-based study was conducted to assess the usefulness of the developed algorithm (identifying episodes of hospitalizations) in 707,916 patients aged 65 years or older, who had received at least one dispensed prescription for Aspirin, Acetaminophen, NSAIDs, or COX-2 inhibitors between April 1, 1999 and December 31, 2002. All medical and pharmaceutical claims, as well as patients' demographic records were obtained from the RAMQ databases. For the same study period, hospital discharge summary data for these patients were also obtained from MED-ECHO.

### 5.2.1 Population Characteristics

Within the initial sample and the use of RAMQ medical services claims database, only patients who had at least one inhospital medical claim were identified between April 1, 1999 and March 31, 2002. The observation period was limited to March 31<sup>st</sup>, since MED-ECHO data are being collected by the same fiscal year—April 1<sup>st</sup> to March 31<sup>st</sup> of each year. It was found that 164,594 individuals were admitted to Quebec hospitals at least once during the study period. According to the RAMQ medical services database, a total of 12,038,887 inpatient medical services were provided to these patients.

In this population, approximately 48% of the patients were males, where the mean age at the time of the entry into the cohort was found to be  $76 \pm 6$  years (mean  $\pm$  SD) with a median of 76 years. Distributions of age and sex were almost similar in the group of individuals with inhospital service claims.

By applying the algorithm with the gap of 7 days on the validation dataset received from the RAMQ medical claims database, inhospital length of stay (LOS) and number of hospitalizations were estimated for the period between April 1<sup>st</sup>, 1999 and March 31<sup>st</sup>, 2002. The parameter estimates derived from this new dataset were then compared to that of observed in MED-ECHO.

The total number of hospitalizations for the target population as identified by the gold standard (MED-ECHO database) was found to be 320,168 hospitalizations for 171,306 patients with an average of  $1.9 \pm 1.5$  [ $\pm$  SD] hospitalizations per individual; and an average length of stay of  $13.7 \pm 18.7$  [ $\pm$  SD] days. However, after applying the algorithm with the gap of 7 days on the validation dataset received from the RAMQ medical claims database, the total number of hospitalizations was estimated to be 310,594 inhospital episodes for 164,594 individuals.

From hospitalizations identified using the RAMQ medical claims database and the use of a deterministic automated matching algorithm, 293,456 hospitalizations were found to be corresponding to that of MED-ECHO (approximately 91%); whereas, 17,138 hospitalizations had no correspondent in MED-ECHO database (missing). Among those matched, 237,235 hospitalizations (77.9%) were estimated with exact lengths of stay; however, number of underestimations and overestimations were found to be 56,221 (19.2%) and 8,651 (2.9%), respectively.

### 5.2.2 Central Tendency

Table 5.5 shows the central tendency (i.e., mean, median, etc.) of the inhospital LOS and number of hospitalizations obtained from the RAMQ medical

claims using the algorithm with the 7 days of gap compared to that of observed in the gold standard.

**Table 5.5** – Estimated population parameters in the RAMQ database compared to values observed in Gold Standard (regardless of admission diagnosis).

	RAMQ (Gap = 7 days) n = 310,594	MED-ECHO (Gold Standard) n= 320,168	P value *
<b>Length of Stay</b>			
Mean	12.9	13.7	< 0.001
Range	1 – 727	1 – 796	
Geometric Mean	8.1	8.1	0.7376
Quantile:			
25% Q1	4.0	4.0	
50% Median	8.0	8.0	0.8954
75% Q3	15.0	16.0	
<b>No. of Hospitalizations</b>			
Mean	2.0	1.9	0.7618
Range	1 – 41	1 – 38	
Geometric Mean	1.0	1.0	0.3456
Quantile:			
25% Q1	1.0	1.0	
50% Median	1.0	1.0	0.7506
75% Q3	2.0	2.0	
* : P value is estimated based on two-sided t-test, however, it is estimated based on median two-sample test for 50% Median estimates. n: Number of hospitalizations			

The appropriateness of various standard measures of central tendency for analyzing in-hospital LOS and number of hospitalizations in our population was also explored. The mean estimated LOS was 12.9 days with the median and geometric mean of 8.0 and 8.1 days in RAMQ, respectively; whereas the average LOS observed

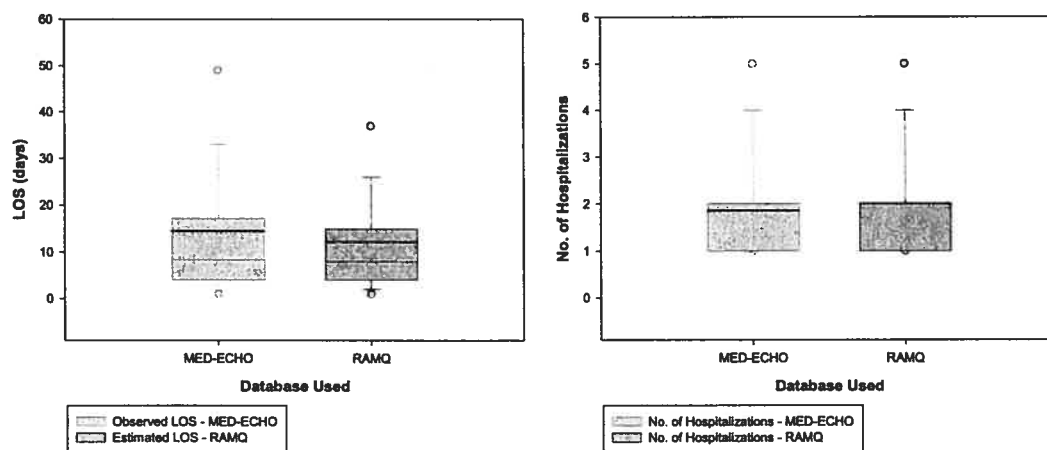


in the MED-ECHO file was 13.7 days with median and geometric mean of 8.0 and 8.1 days. The average number of hospitalizations was estimated to be 2.0 versus 1.9 for the entire study sample in the RAMQ and MED-ECHO databases, respectively. Although the mean LOS and number of hospitalizations were slightly different in the two databases, it was shown that there was no statistically significant difference between medians and geometric means.

The mean inhospital length of stay and the average number of hospitalizations in our target population were significantly higher than the medians, reflecting the rightward skewness. That is, patients with extended stays heavily influenced mean length of stay. Moreover, the geometric means were lower than the arithmetic means for both parameters; they tended to track the median values. The geometric mean for this analysis was computed as the mean of the natural logarithm transformation. Log transformation decreased the skewness of the frequency distributions and affects extremely high values, where only a few patients per length of stay or number of hospitalizations were existed at extreme values.

Figure 5.12 demonstrates the parameter estimates in the RAMQ data compared to the observed values in the gold standard. In the box plots, ranges or distribution characteristics of the variables are plotted separately for each database. The central tendency, midrange, as well as 5%-95% data range (whiskers) computed for each group are presented in this figure. Because of the long-tailed distribution (rightward

skewness) and the shrinkage of the box plots, points beyond the range of 5%-95% are not shown.

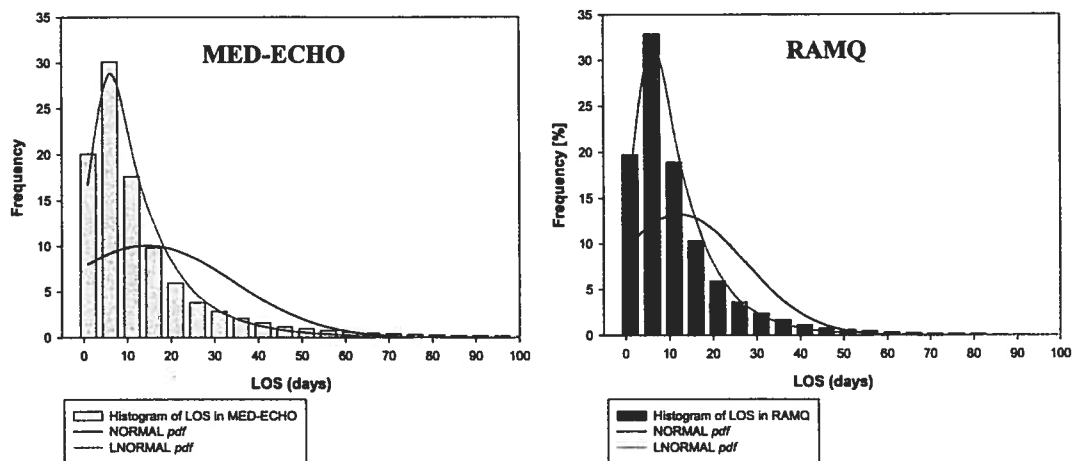


**Figure 5.12** – Estimated LOS and N° of hospitalizations compared to MED-ECHO database (gold standard) regardless of admission diagnosis.

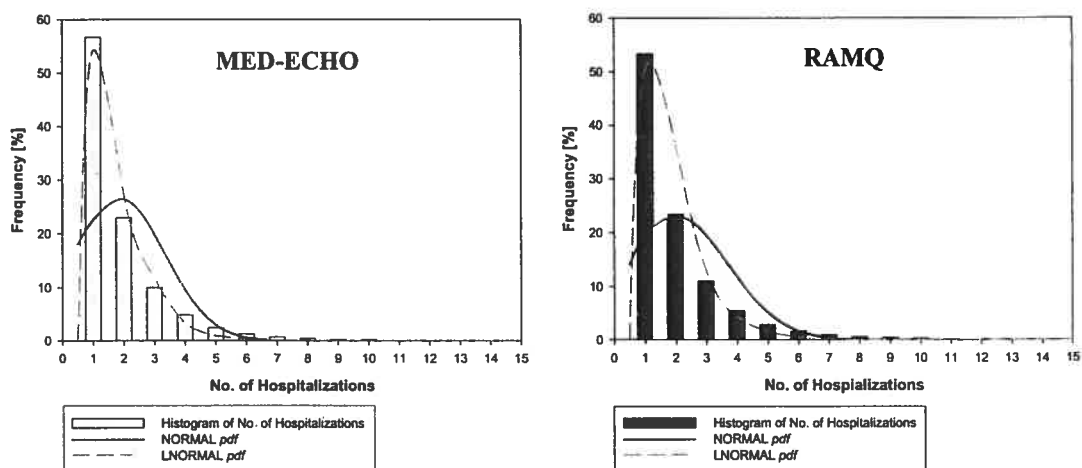
Overall, the inhospital LOS was slightly underestimated in the validation dataset, whereas the average number of hospitalizations was correctly estimated. The results obtained from this analysis show that the algorithm with the gap of 7 days can be used to correctly assess episodes of hospitalizations in the RAMQ medical claims database.

Figure 5.13 and Figure 5.14 show histogram distributions and the probability distribution functions (pdf) of LOS and number of hospitalizations in the RAMQ medical claims database compared to that of observed in MED-ECHO file as the gold standard, respectively. As shown in these figures, LOS and number of hospitalizations

can be properly approximated using lognormal distributions. This is a very important finding since the developed algorithm can be used as a predictive model to identify potential factors (covariates) affecting episodes of hospitalizations in future studies.

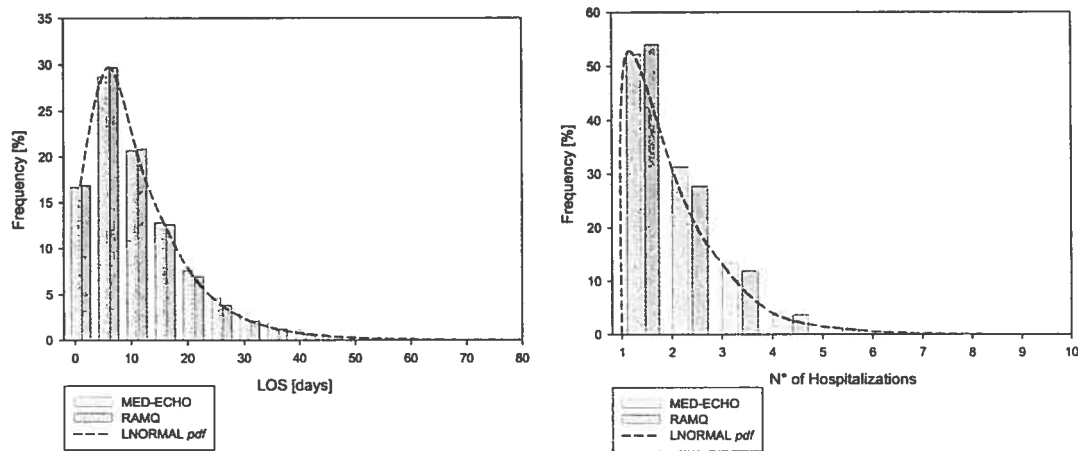


**Figure 5.13** – Histogram and the *pdf* of LOS in MED-ECHO and RAMQ databases.



**Figure 5.14** – Histogram and the *pdf* of N° of Hospitalizations in both databases.

Figure 5.15 demonstrates the comparison of parameter estimates (LOS and number of hospitalizations) obtained from the RAMQ medical claims database to that of observed in MED-ECHO. Visual comparison of the parameter estimates obtained from the RAMQ medical claims database with the observed values in MED-ECHO also indicated a slight underestimation of the in-hospital LOS using the RAMQ file.



**Figure 5.15** – Comparison of the estimated LOS and N° of hospitalizations with the gold standard regardless of admission diagnosis; both pdfs follow lognormal distributions.

For the validation database, service interruptions (the gap) in the RAMQ medical claims database and the intervals between corresponding hospitalizations in the MED-ECHO database were also computed for the entire study period. This was done to reassure the appropriateness of the algorithm with the gap of 7 days.

Average length of gap in the RAMQ file was  $73.9 \pm 149.5$  [ $\pm$ SD] days with the median and mode equal to 7.0 and 2.0 days, respectively; whereas average interval between hospitalizations in MED-ECHO appeared to be  $169.3 \pm 190.9$  [ $\pm$ SD] days with the median and mode of 94.0 and 7.0 days, respectively.

Distributions of gaps in the RAMQ and intervals between hospitalizations in MED-ECHO were highly skewed to the right. The median gap in the RAMQ approximately represented the average time by which half of the patients had not received any inhospital services, which could be either service interruptions within a unique hospitalization or the time elapsed between two separate hospitalizations.

Comparing the central tendency measures of the gap in the RAMQ to those of the hospitalization intervals in MED-ECHO indicated that applying the algorithm with the gap of 7 days (to the RAMQ validation dataset) did not result in erroneous estimates (Table 5.5). Less than 5% of the hospitalizations in MED-ECHO were separated with an interval of 8 days; i.e., the proportion of the incorrect overlapping hospitalizations was very small for the entire population in MED-ECHO. Therefore, using the developed algorithm with the gap of 7 days would only result in slightly biased estimations of the LOS and number of hospitalizations in the RAMQ. Figure 5.16 demonstrates the distribution of the gap in the RAMQ medical claims database, whereas Figure 5.17 represents the distribution of hospitalization intervals in MED-ECHO.

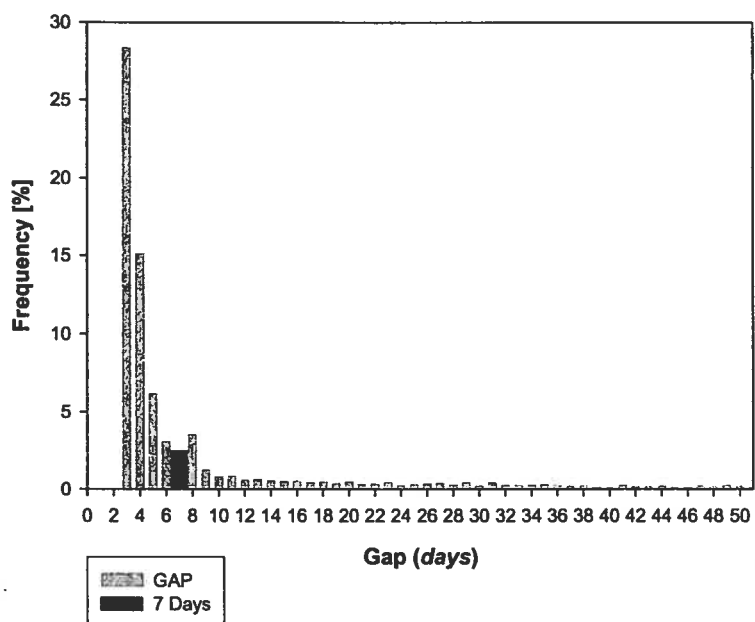


Figure 5.16 – Frequency distribution of gaps in the RAMQ medical claims database.

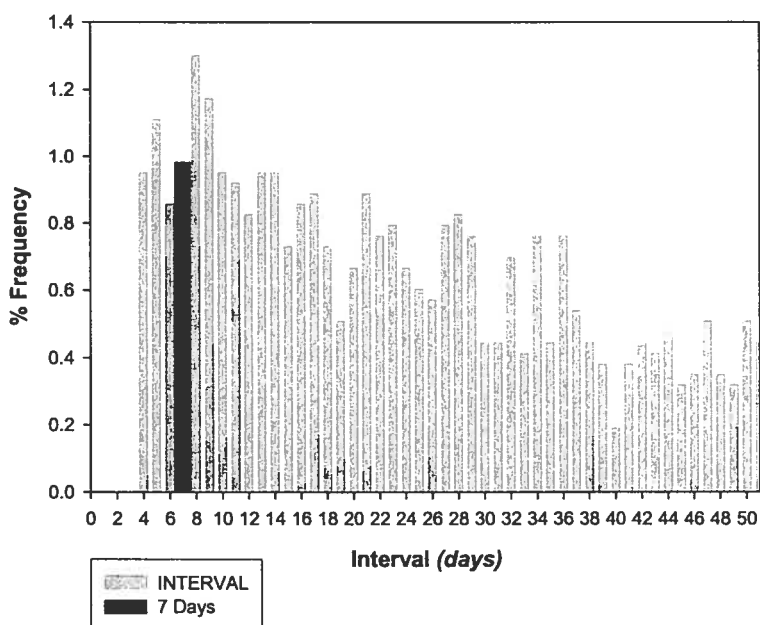


Figure 5.17 – Frequency distribution of hospitalization intervals in MED-ECHO.

### **5.2.3 Hospitalizations for Specific Indications**

The secondary objective of this study was to identify the episodes of hospitalizations for acute myocardial infarction (AMI) and upper gastrointestinal bleeding (UGIB) using the RAMQ medical services database. Based upon case definitions described, in-hospital claims were identified for these two clinical conditions. According to the previous findings of this study, the physician claims were linked to the hospitalizations estimated using the developed algorithm with the gap of 7 days. Estimated parameters from the RAMQ medical claims were compared to those observed in MED-ECHO database in terms of the central tendency and the agreement between the two databases. The results obtained for AMI and UGIB analyses are separately presented in the remaining of this chapter.

#### **5.2.3.1 Acute Myocardial Infarction (AMI)**

The accuracy and validity of using ICD-9 diagnosis code '410', in the RAMQ medical claims database, to identify episodes of hospitalizations for AMI were examined using the MED-ECHO database as the gold standard. This was done using a deterministic automated matching algorithm. Only the patients' first hospitalizations were considered for the agreement analysis (i.e., each individual as a unit of analysis).

#### 5.2.3.1.1 Diagnosis Agreement

The total number of hospitalization episodes for AMI in our target population, as identified by the MED-ECHO database, was found to be 20,848 hospitalizations for 19,076 patients; however, it was estimated to be 25,131 hospitalizations for 23,621 individuals using the RAMQ database. Among patients with an AMI, approximately 55% of individuals were males, where the average age at the index hospitalization was found to be  $75 \pm 6$  [ $\pm$  SD] years for males and  $77 \pm 7$  years for females, respectively.

Compared to MED-ECHO, the sensitivity and the specificity of using ICD-9 diagnosis code '410' as a screening tool to identify first AMI hospitalizations were 81.09% (95%CI: 80.52-81.64%) and 98.82% (95%CI: 98.79-98.84%), respectively. ICD-9 code '410' has high specificity and acceptable sensitivity for detecting the first AMI hospitalizations in the RAMQ claims database.

The efficiency of the RAMQ medical services database for estimating episodes of hospitalizations for AMI was estimated to be 98.34% (95%CI: 98.31-98.37%) when compared to the MED-ECHO database as the gold standard. Efficiency was assessed as the sum of the true positive and true negative divided by the total number of admissions (i.e., the correct classification rate). The kappa statistics (overall agreement) in our analysis was found to be  $\kappa=0.72$  (95%CI: 0.71-0.72), indicating substantial agreement between the two databases for identifying AMI hospitalizations. Since kappa values become low when the prevalence deviates from 50%, we report



kappa coefficients adjusted for prevalence and bias (PABAK). For determination of UGIB hospitalization, the index of agreement was substantial but improved to almost perfect (PABAK = 0.74) after prevalence and bias adjustments. For more details and complete results on the validity of ICD-9 code '410' and the appropriate agreement statistics please refer to Appendix 3, Table 3A.

#### **5.2.3.1.2 Central Tendency**

Inhospital LOS and number of hospitalizations for AMI were then estimated for the period between April 1<sup>st</sup>, 1999 and March 31<sup>st</sup>, 2002. Table 5.6 shows the central tendency (i.e., mean, median, etc.) of the inpatient LOS and number of hospitalizations for AMI obtained from the RAMQ medical claims using the optimal algorithm (gap = 7 days) and ICD-9 diagnosis code '410' as a screening tool for AMI hospitalization.

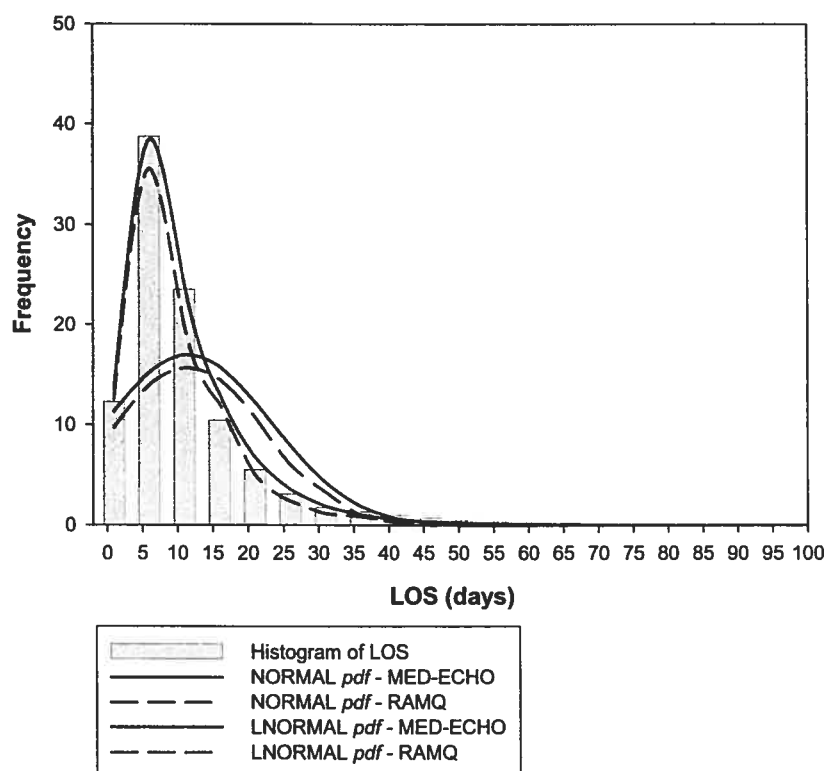
The mean estimated LOS for AMI was 8.7 days with the median and geometric mean of 7.0 and 7.2 days in RAMQ, respectively; whereas the average LOS observed in the MED-ECHO database was found to be 11.5 days with median and geometric mean of 8.0 and 8.0 days. The average number of hospitalizations was estimated to be 1.1 versus 1.0 for the entire study sample in the RAMQ and MED-ECHO databases, respectively. Although the average LOS was significantly different

in the two databases, it was shown that the two have slight differences in median LOS or the geometric mean. Since the majority of patients were hospitalized only once during the study period (i.e., approximately 92% in MED-ECHO and 94% in RAMQ), the average number of hospitalizations in the RAMQ database did not differ from the one observed in MED-ECHO. The geometric means of all parameters tended to track the median values.

**Table 5.6** – Estimated population parameters for AMI in the RAMQ database compared to the values observed in MED-ECHO.

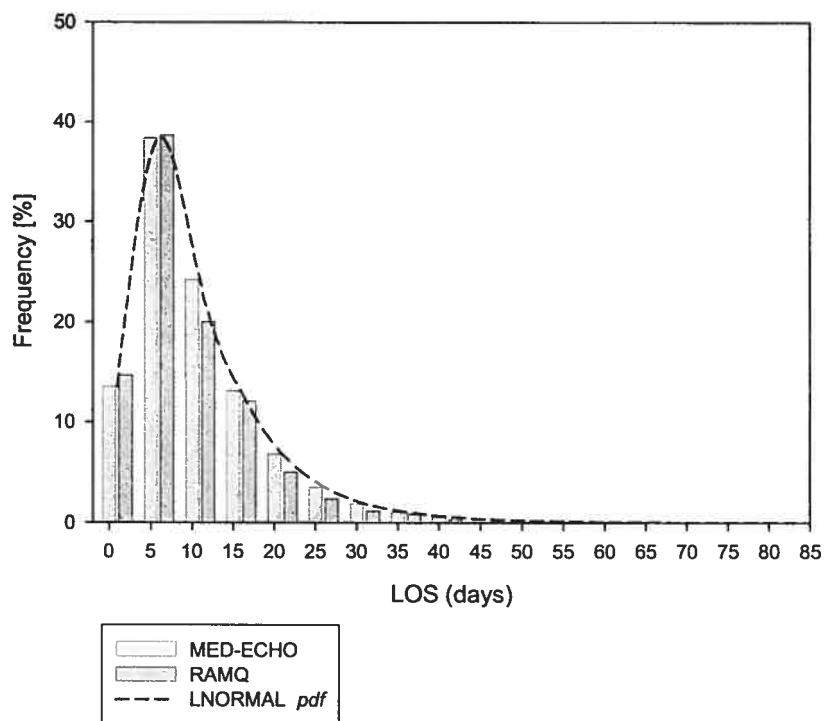
	RAMQ (Gap = 7 days) n = 25,131	MED-ECHO (Gold Standard) n= 20,848	P value *
<b>Length of Stay</b>			
Mean	8.7	11.5	< 0.001
Range	1 – 92	1 – 159	
Geometric Mean	7.2	8.0	< 0.001
Quantile:			
25% Q1	4.0	5.0	
50% Median	7.0	8.0	< 0.001
75% Q3	11.0	14.0	
<b>No. of Hospitalizations</b>			
Mean	1.1	1.0	> 0.05
Range	1 – 9	1 – 7	
Geometric Mean	1.0	1.0	> 0.05
Quantile:			
25% Q1	1.0	1.0	
50% Median	1.0	1.0	> 0.05
75% Q3	1.0	1.0	
* : P value is estimated based on two-sided t-test, however, it is estimated based on median two-sample test for 50% Median estimates.			
n: Number of hospitalizations.			

Overall, the in-hospital LOS for AMI was underestimated in the validation dataset, whereas the average number of hospitalizations was correctly estimated. Figure 5.18 demonstrates histogram distribution and the probability distribution functions (pdf) of the LOS for AMI in the RAMQ medical claims database compared to that of observed in MED-ECHO; since the majority of subjects had only one hospitalization during the study period, no graphical presentation is therefore provided for the number of hospitalizations distribution.



**Figure 5.18** – Histogram and the *pdf* distribution of LOS in MED-ECHO and RAMQ.

The comparison of estimated LOS from the RAMQ medical claims database with that of observed in MED-ECHO is shown in Figure 5.19. Visual comparison of the estimated LOS in RAMQ with the gold standard indicated an apparent underestimation of this parameter using the RAMQ claims database for the entire range.



**Figure 5.19** – Comparison of the estimated LOS in the RAMQ with MED-ECHO as the gold standard; LOS follows a lognormal distribution.

### **5.2.3.2 Upper Gastrointestinal Bleeding (UGIB)**

The accuracy and validity of using endoscopy procedural codes and/or ICD-9 diagnostic codes of UGIB in the RAMQ medical claims database, to identify episodes of hospitalizations for UGIB was examined using the MED-ECHO database as the gold standard. This was performed by means of a deterministic automated matching algorithm using unique patients' identification numbers. Only the patients' first hospitalizations were again considered for the agreement analysis (i.e., each individual as a unit of analysis).

#### **5.2.3.2.1 Diagnosis Agreement**

The total number of hospitalization episodes for UGIB in our target population, as identified by the MED-ECHO database, was found to be 5,623 hospitalizations for 5,280 patients. Using different models (M1-M4), however, resulted in a variety of estimates of the number of hospitalizations, as well as different number of individuals, for UGIB in the RAMQ physician claims database.

As previously described in Chapter 4 (Methodology), potential hospitalizations for UGIB in model M1 were identified if there was at least one physician claim for upper gastrointestinal endoscopy within the first 24 hours of the admission date.

Model M2 used a combination of ICD-9 codes for UGIB, including 531.X, 532.X, 533.X, 534.X, and 578.X. In model M3, hospitalizations that contained a code for gastrointestinal hemorrhage (ICD-9 578.X) and had one or more codes for specific gastrointestinal tract disorders (such as neoplasm, esophageal varices, etc.) were excluded. Further restriction to model M3 in which only hospitalizations that included a claim for an upper gastrointestinal endoscopy within the first 24 hours of the admission date resulted in model M4.

Validity was estimated using various statistics for each model as the proportion of responses for the hospitalization in which the observed MED-ECHO and the estimated RAMQ hospitalizations were in agreement. The sensitivity, specificity, positive and negative predictive values, and efficiency of using each model as a screening tool to identify first hospitalizations for upper gastrointestinal bleeding or perforation are shown in Table 5.7.

As shown in Table 5.7, model M1 had a high specificity (99.2%) and fair sensitivity (61.9%). Nevertheless, model M2 with a combination of ICD-9 codes has a higher specificity (99.3%) and an acceptable level of sensitivity (73.7%) for detecting UGIB hospitalizations. For model M3, the specificity increased (99.4%) marginally, while sensitivity declined (69.7%). Ascertainment of having an upper gastrointestinal endoscopy within the first 24 hours of the admission date (model M4) further increased the specificity (99.8%); however, the sensitivity was rather lower (65.1%).

**Table 5.7:** Selected agreement statistics comparing different models for UGIB.

<b>Model</b>	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>
N° of Individuals	8899	9034	7822	6130
<b>Statistics (%) (95% CI)</b>				
Sensitivity	61.9 (60.6-63.2)	73.7 (72.5-74.9)	69.7 (68.4-70.9)	65.1 (63.8-66.4)
Specificity	99.2 (99.1-99.2)	99.3 (99.2-99.3)	99.4 (99.3-99.4)	99.8 (99.7-99.8)
PPV	36.7 (35.7-37.7)	43.1 (42.0-44.1)	47.0 (45.9-48.1)	66.9 (65.6-68.2)
NPV	99.7 (99.7-99.7)	99.8 (99.7-99.8)	99.8 (99.7-99.8)	99.7 (99.7-99.8)
Efficiency	98.9 (98.8-99.0)	99.1 (99.0-99.1)	99.2 (99.1-99.2)	99.5 (99.4-99.5)
⇒ <b>M1:</b> A physician claim for upper gastrointestinal endoscopy within 24 hrs. ⇒ <b>M2:</b> Specific and non-specific diagnostic codes for gastrointestinal hemorrhage. ⇒ <b>M3:</b> Specific diagnostic codes for gastrointestinal hemorrhage. ⇒ <b>M4:</b> Combination of models M1 and M3. ⇒ <b>PPV :</b> (predictive value of positive test); <b>NPV :</b> (predictive value of negative test); <b>95%CI</b> (95% confidence interval)				

The positive predictive value of model M4 was higher (66.9%) than other models particularly model M1 (36.7%), where the only criterion for identification of UGIB hospitalizations was a physician claim bearing a procedural code for upper gastrointestinal endoscopy. On the other hand, the negative predictive values were high, relatively similar for all models. Overall, the efficiency was improved when screening for UGIB hospitalizations was implemented using the model M4 (99.5%).

**Table 5.8:** Reliability statistics comparing different models for UGIB.

<b>Model</b>	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>
N° of Individuals	8899	9034	7822	5138
<b>Statistics</b>				
Kappa	0.46 (0.45-0.47)	0.54 (0.53-0.55)	0.56 (0.55-0.57)	0.66 (0.65-0.67)
BI	-0.0051	-0.0053	-0.0036	0.0002
PAI	0.9800	0.9798	0.9815	0.9853
PABAK	0.9784	0.9815	0.9838	0.9900
LR+	77.23 (74.68-79.87)	100.6 (97.5-103.9)	118.1 (114.0-122.3)	268.9 (255.4-283.1)
LR-	0.38 (0.37-0.40)	0.27 (0.25-0.28)	0.31 (0.29-0.32)	0.35 (0.34-0.36)
Odds Ratio (OR)	201	379	387	768
<i>P in</i> *	10.24	21.07	19.37	26.44
<i>P out</i> *	2.50	3.61	3.16	2.81

⇒ **M1**: A physician claim for upper gastrointestinal endoscopy within 24 hrs.

⇒ **M2**: Specific and non-specific diagnostic codes for gastrointestinal hemorrhage.

⇒ **M3**: Specific diagnostic codes for gastrointestinal hemorrhage.

⇒ **M4**: Combination of models M1 and M3.

⇒ **BI**: Byrt's bias index; **PAI**: Byrt's prevalence asymmetry index; **PABAK**: prevalence and bias adjusted kappa; **LR+**: positive likelihood ratio; **LR-**: negative likelihood ratio.

⇒ \* Rule-in (*P in*) and Rule-out (*P out*) potentials which were computed based on the Kullback-Leibler distance.



Reliability analyses for different models identifying UGIB hospitalizations were also performed; the results of these analyses are presented in Table 5.8. For identification of UGIB hospitalizations, inter-observer agreement was only moderate ( $\kappa = 0.46$ ) for pairwise  $\kappa$  for model M1; however, kappa statistics improved with the specificity of the model used. The most reliable model was found to be model M4 with the highest kappa coefficient ( $\kappa = 0.66$ ), which indicated substantial agreement between the two data sources. Moreover, kappa statistics remarkably improved to ‘*almost perfect*’ after prevalence and bias adjustment (e.g., prevalence and bias adjusted kappa was found to be 0.99 for model M4).

As shown in Table 5.8, summarizing positive and negative likelihood ratios, the likelihood ratio test results and the overall odds ratio were higher for model M4 (OR=768) compared to the other models. The rule-in and rule-out potentials, which were calculated using the Kullback-Leibler distance (before-test potentials of ruling-in and -out), were different for various models as well (Table 5.8). The rule-in potential (*P in*) for the model M4 was 26.44, and 21.07 for the model M2; the rule-out potentials (*P out*) were 2.81 and 3.61, respectively. In other words, for model M4, an individual with UGIB hospitalization would be 26.44 times greater, on average, of being ruled in as having a UGIB hospitalization. While for the same model, a person with no hospitalization would become 2.81 times more likely to be marked as having no hospitalization. Furthermore, it was apparent that the most specific algorithm had higher rule-in potential, while the most sensitive one performed better at ruling out

UGIB hospitalizations. For complete list of agreement statistics on the validity of ICD-9 codes for UGIB on RAMQ claims please refer to Appendix 4, Table 4A-4D.

#### **5.2.3.2.2 Central Tendency**

Among patients with an UGIB, approximately 54% of individuals were males, where the average age at the index hospitalization was found to be  $75 \pm 6$  [ $\pm$  SD] years for males and  $78 \pm 6$  years for females, respectively.

Inhospital LOS and number of hospitalizations for UGIB were estimated for the period between April 1<sup>st</sup>, 1999 and March 31<sup>st</sup>, 2002. Table 5.9 summarizes the results of the central tendency (i.e., mean, median, etc.) for inpatient LOS and number of hospitalizations obtained from the RAMQ medical claims database using the most sensitive (M2) and the most specific algorithm (M4). The comparison of the parameter estimates to those observed in MED-ECHO is also presented (Table 5.9).

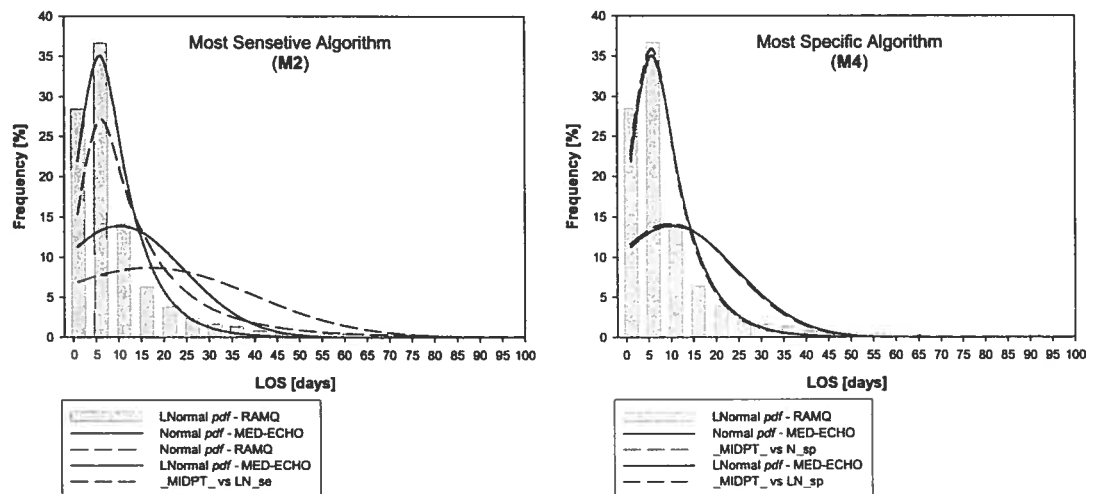
As observed in the gold standard (MED-ECHO database), mean inpatient LOS for UGIB was 10.3 days with an equal median and geometric mean of 6.0 days, respectively; while the average number of hospitalizations was found to be 1.2, with median and geometric mean of 1.0. For the most sensitive algorithm (model M2), the mean estimated LOS for UGIB was 16.9 days with median and geometric mean of 9.0 and 8.9 days, respectively; whereas, the average LOS estimated by the most specific algorithm (model M4) was found to be 9.9 days with median and geometric mean of

6.0 days. The average number of hospitalizations was estimated to be 1.1 versus 1.2 for the entire study sample using the RAMQ and MED-ECHO databases, respectively. As shown in Table 5.9, the best estimates for the in-hospital LOS and number of hospitalizations were derived when model M4 (the most specific algorithm) was used. In fact, by applying model M4, there were only slight differences between the estimated parameters in the RAMQ and MED-ECHO databases. The geometric means of all parameters tended to track the median values in the RAMQ medical claims using model M4, as well as the MED-ECHO database.

**Table 5.9** – Estimated population parameters for UGIB in the RAMQ database compared to the parameters observed in MED-ECHO.

	RAMQ (M2) n = 10,013	RAMQ (M4) n = 5,665	MED-ECHO (Gold Standard) n = 6,965
<b>Length of Stay</b>			
Mean	16.9	9.9	10.3
Range	1 – 324	1 – 229	1 – 189
Geometric Mean	9.0	6.0	6.0
Quantile:			
25% Q1	4.0	3.0	3.0
50% Median	9.0	6.0	6.0
75% Q3	12.0	11.00	11.0
<b>No. of Hospitalizations</b>			
Mean	1.1	1.1	1.2
Range	1 – 7	1 – 8	1 – 11
Geometric Mean	1.0	1.0	1.0
Quantile:			
25% Q1	1.0	1.0	1.0
50% Median	1.0	1.0	1.0
75% Q3	1.0	1.0	1.0
n: Number of hospitalizations.			

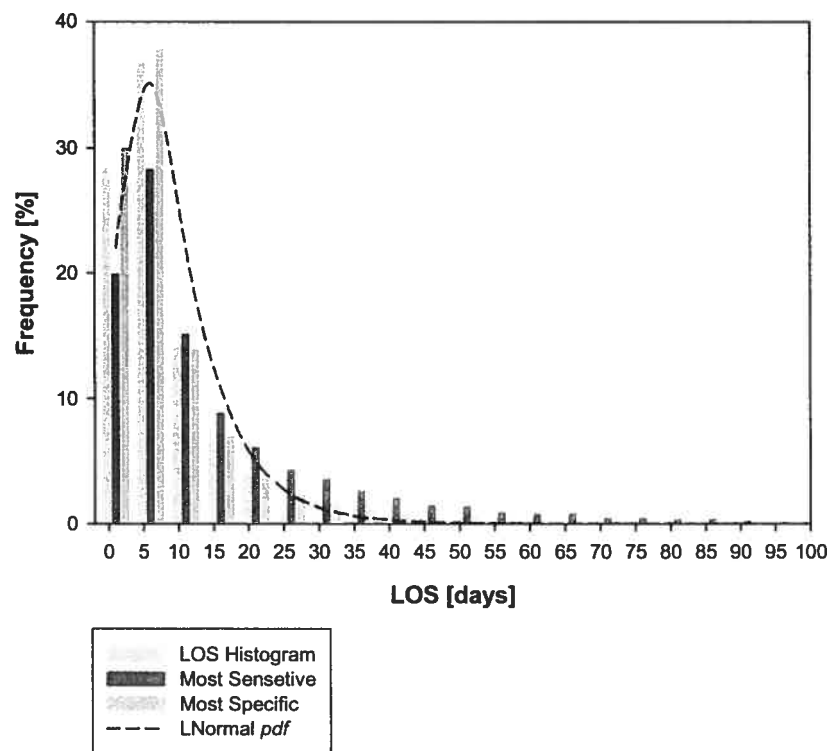
Overall, the in-hospital LOS for UGIB was slightly underestimated in the validation dataset, whereas the average number of hospitalizations was correctly estimated. Figure 5.20 demonstrates histogram distribution and the probability distribution functions (pdf) of the LOS for UGIB in the RAMQ medical claims database compared to that of observed in MED-ECHO; since the majority of subjects had only one hospitalization (*average*  $\approx 1.0$ ) during the study period, no graphical presentation is therefore provided for the distribution of number of hospitalizations.



**Figure 5.20** – Histogram and the *pdf* distributions of LOS in MED-ECHO and RAMQ for models M2 (most sensitive) and M4 (most specific).

For a better comparison of estimated LOS from the RAMQ medical claims database using models M2 and M4 with that of MED-ECHO, their frequency

distributions are simultaneously shown in Figure 5.21. Visual comparison of the estimated parameter in the RAMQ claims data with that of the gold standard indicated the superiority of model M4 in estimating in-hospital LOS for UGIB. As shown, LOS can be estimated with a great accuracy using the RAMQ medical claims database.



**Figure 5.21** – Comparison of the estimated LOS in the RAMQ with MED-ECHO as the gold standard; distribution of LOS can be approximated using a lognormal distribution.

# ***Discussion***

---

# **6**

Health care administrative databases are being used in many epidemiologic and outcomes research studies, including studies of resource consumption, physician practice patterns, and outcome assessment. In fact, using these data sources for such studies has become increasingly important in the health care setting [Wang FL, 2001; Sorensen HT, 1996]. Previous studies have evaluated the accuracy of the data elements within several Canadian administrative databases by assessing the agreement with data reabstracted from charts or other electronic databases. These studies showed that demographic information, primary diagnoses, and some procedural codes were accurately recorded [Newfoundland Department of Health, 1995; Delfino RJ, 1993; The Doctor's Hospital, 1992; Ontario Hospital Association, 1991]. Nevertheless, comorbidities demonstrated poor agreement [Newfoundland Department of Health, 1995; Malenka DJ, 1994; Delfino RJ, 1993; Ontario Hospital Association, 1991].

Hospital discharge summary data have been the primary source of information for these studies, particularly for outcome assessment. The validity of such data from different sources has been previously assessed for a number of clinical conditions in which they appeared to be accurate [Pervez H, 2003; Austin PC, 2002; Cooper GS, 2000; Jollis JG, 1993; Tennis P, 1993]. More specifically, in the Canadian province of Quebec, hospital discharge summary (MED-ECHO) was found to be both accurate

and reliable [Levy AR, 1999; Levy AR, 1995]. However, the validation of outcome assessment in hospitalized patients using the RAMQ physician claims data is a rather unexplored field. Usually, the reluctance to use physician claims data in outcome research is likely related to both technical challenges in using the data and skepticism about the utility of the diagnostic information recorded [Tamblyn R, 2000; Mitchell JB, 1994; Mitchell JB, 1992]. Assessing the accuracy and validity of this information would not only shed light on the reliability of the RAMQ medical services database as a data source for research, but also it would assist researchers in the data analysis and the interpretation (statistical inference) of the results.

While the quality of data for research is one of the most important concerns of all health-related studies, little attention is being paid to the validity of the raw data being analyzed. Irrespective of the data source, the accountability of a study's conclusions rests on the accuracy and validity of the data being used. For instance, in an outcome research study, the collection of data must cover the primary care setting including inhospital records, which should strongly depend on a solid information base. This becomes even more important if a patient is frequently seen by the physician and may further require inhospital admission.

The accuracy and reliability of inpatient records in the RAMQ medical services database have never been assessed for specific clinical conditions, such as COPD, AMI, and UGIB; no previously published literature on these subject matters were identified. This dissertation examined the accuracy and reliability of the RAMQ

medical services data for identifying the cause, duration, and number of hospitalizations for these clinical conditions in Quebec, Canada.

## 6.1 Hospitalization Episodes

Using the RAMQ medical services database, episodes of hospitalizations were estimated by applying an analytical algorithm with a model variable controlling presumed service interruption(s) (*the gap*) within a patient's unique hospitalization. The parameter estimates were then compared to those observed in the gold standard, MED-ECHO, through a deterministic linkage algorithm. Comparison of the parameter estimates (analyses of central tendency, predictive performance, and agreement), was performed to select the optimal algorithm with the most accurate estimates for the actual observations in the gold standard. The results of this study demonstrated that the algorithm with the gap of 7 days can yield relatively accurate estimates for the length of stay and number of hospitalizations.

Estimated parameters, such as hospitalization episodes, from large, automated databases can also be useful to study the natural history of respiratory diseases in the community, given that the validity of information is demonstrated. Derived from the estimated hospitalizations using the optimal gap, the prevalence of COPD was found to be age-dependent which is in total agreement with the previous study by the



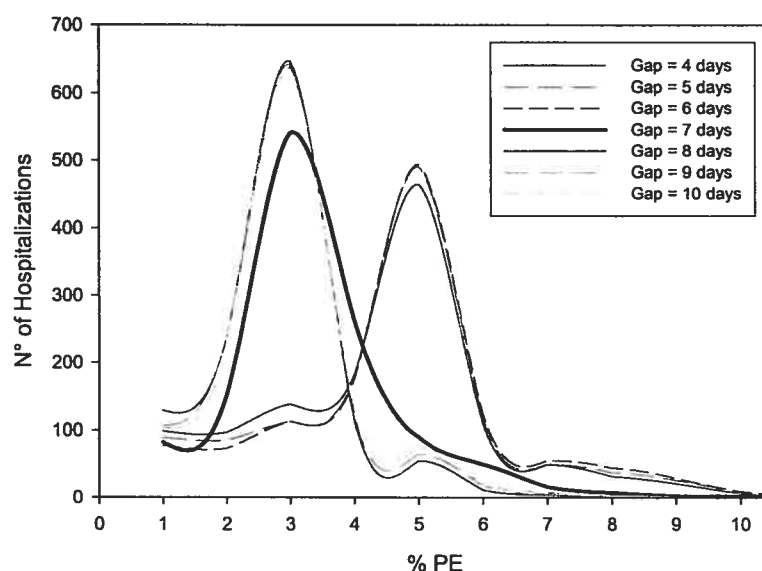
Canadian National Institutes of Health [Petty TL, 1997; National Institutes of Health, 1996]. COPD symptoms are more likely to start at the earlier age, since the majority of COPD cases are usually caused by long term smoking.

The remainder of this chapter will first provide evidence to justify the utilization of gap of 7 days in both development and validation data sets. The appropriateness of various standard measures of central tendency for analyzing in-hospital LOS and number of hospitalizations will then be considered. The application of the developed algorithm to identify hospitalizations for specific diagnoses (i.e., COPD, AMI, and UGIB hospitalizations) will be presented. Lastly, diagnostic agreement between RAMQ medical services claims and MED-ECHO discharge summary data will be discussed.

### 6.1.1 The Optimal Gap

The predictive performance analyses (i.e., bias, accuracy, and precision) on the parameter estimates strongly pointed at 7 days as the optimal gap for identifying hospitalization episodes in our study. According to the results of these analyses, this algorithm ( $gap = 7 \text{ days}$ ) can provide the most accurate and precise estimates with a relatively small bias for hospitalization episodes.

Inferences were not limited to the point estimates of the predictive performance only; distributions of different measures were also considered to provide help for the selection of the optimal algorithm. For instance, Figure 6.1 demonstrates the distribution of percent predictive error (%PE) of the estimates for selected algorithms. As shown in this figure, the algorithm with the gap of 7 days tends to narrow down the extent of the variability existed on the prediction estimates.



**Figure 6.1** – Comparison of percent predictive errors (%PE) using different gaps.

Furthermore, the gap (i.e., interruption between services that a patient might experience during a unique hospitalization) between the claims in the RAMQ medical services database and the interval between hospitalizations in the MED-ECHO

database were measured in our studies. In both data sets (development and validation), graphical presentation of the distributions of gaps and intervals revealed a weekly pattern. Using this pattern, one can simply measure the minimum and maximum rates of the service interruption in the claims data at different time points.

The most frequent interval between hospitalizations in MED-ECHO was found to be 8 days which did not overlap with the optimal gap in the developed algorithm. Hospitalizations that were apart with an interval less than 8 days comprise a small fraction in MED-ECHO database (less than 5%). Therefore, only a small fraction of MED-ECHO hospitalizations would be affected by using the gap of 7 days, by which in-hospital LOS was prone to overestimation.

Using the algorithm with an optimal gap of 7 days, the matching procedures were quite successful in both analyses (i.e., training and validation data sets). The correct matching rate was 81% in the training data set for the cohort of COPD patients, using unique patient identifier and dates of admission and discharge. This rate was even higher (91%) when the developed algorithm was applied to the validation data set. The improvement in the rate of successful matching was expected since no restrictions (i.e., no specific selection criteria on diagnosis) were applied to the patients in the validation data set.

In the linkage procedures, failure in matching was largely caused by missing or incomplete information on the variables used, such as unique patients' identifiers

and dates of admission and discharge. As with other linkage methods, the success of deterministic linkage depends largely on the completeness and accuracy of the information in the files to be linked and an appropriate combination of matching variables.

As suggested by the successful matching rates for both data sets received from the RAMQ, the quality and completeness of RAMQ medical services database regarding in-hospital claims were found to be relatively high; this provides promise for future studies using the RAMQ medical services database for estimating episodes of hospitalizations, including LOS and number of hospitalizations.

Overall, results and graphical presentations of the estimated parameters clearly showed that the developed algorithm with the gap of 7 days can potentially be used as an analytical tool to measure episodes of hospitalizations from the RAMQ medical services database.

### 6.1.2 Appropriateness of Central Tendency

The developed algorithm generated realistic estimates of in-hospital episodes in the elderly, which were in total agreement with observed values in the gold standard. When estimates of in-hospital stays and number of hospitalizations obtained from the RAMQ claims data were compared to those observed in MED-ECHO, the average in-hospital LOS was slightly lower for the RAMQ claim-based estimates, whereas the median and the geometric means were roughly the same. The average, as well as the median number of hospitalizations was strongly similar in both data sets (i.e., the RAMQ and the gold standard). The LOS finding is in total agreement with the assumption that claims data consistently underestimate the degree of in-hospital morbidities, performance, and utilization indicators [Halfon P and Egli Y, 2001; Kashner TM, 1998; Fleming C, 1992; Roos LL, 1991; Potvin L and Champagne F, 1986].

Hospital morbidity indices, such as LOS and number of hospitalizations, are important measures of hospital activity and health care utilization; they have implications for service planning, resource allocations, bed utilization, and determinants of hospitalization costs. Hospital morbidity data can provide managers with much information on the utilization of hospital resources, providing they pay attention to the characteristics of the data. Unfortunately, the empirical distributions of LOS and number of hospitalizations are often positively skewed and vary

considerably across diagnoses [Marazzi A, 1998]. The skewness of these variables often poses a problem for statistical analysis.

In our analyses, the skewed distribution patterns required that various descriptors of central tendency be evaluated to determine which best described the "typical" length of stay. The "typical" length of stay is important, since it provides a basis from which to evaluate changes in routine practice unencumbered by outliers [Sperry S and Birdsall C, 1994]. The frequency distributions were usually non-gaussian or not "normal" (bell shaped) and zero-bound (stays could never be less than zero days). Skewness to the right was commonplace, but not unexpected, particularly in the COPD cohort given the nature of the patient population, prolonged rather than short episodes of stay were more common.

The distributions usually included two subpopulations of interest: The portion with the majority of observations ("body") and the "tail". Each provided different information. The body described "typical" behavior, i.e., how long most patients in a particular diagnostic group stayed in the hospital. The tail provided information on the exceptions (outliers), those patients with prolonged stays. Each part of the distribution is important for clinical and administrative reasons. Therefore, two major issues were examined: How to best describe the "typical" length of stay (the central tendency), and how best to identify outliers.

The first step in examining these issues was visual analysis of frequency distributions and cumulative frequencies. Such graphs easily revealed the characteristics of the data (i.e., normality, skewness etc.), helped determine which measure of central tendency was optimal, and also helped identify outliers and other unusual characteristics.

#### **6.1.2.1 Typical Measure**

Unlike the arithmetic mean, the median (or geometric mean) successfully described the "typical" length of stay. The median is a well understood and routinely used summary statistics for variables whose distributions are skewed or heavy-tailed. As an estimate of central tendency, the median is a resistant measure that is not as greatly affected by outliers (i.e., those with prolonged LOS) as is the arithmetic mean. As expected in a rightwardly skewed distribution, both median and geometric mean were smaller than the mean [Ratcliff R, 1993; Ottenbacher KJ, 1993; Maxfield M, 1988]. Arithmetic means were heavily influenced by outliers and were not appropriate for describing the "typical" length of stay. Consequently, in epidemiologic studies and outcomes assessment the median value seems to be more promising. However, for financial and other planning purposes, the mean is useful, as it accounts for patients

with extended stays [Finlayson G, 2002; Birnbaum HG, 1999; Clewer A and Perkins D, 1998; Edwards WH, Sr., 1996].

The median has some limitations, however. In very long tailed distributions, it may not accurately represent the central tendency [Ratcliff R, 1993]. In our analyses, the geometric means appeared to be alternatives to the median.

#### **6.1.2.2 Outliers**

Identifying outliers is an area of much interest in both the medical and statistical literature [Tu JV, 1994; Chow SC and Tse SK, 1990; Dandona L, 1989]. From an administrative standpoint, it is important to identify outliers so that the reasons for their prolonged stays can be determined and steps instituted to prevent future prolonged stays. Investigation of the clinical course of outliers may uncover unacceptable complication rates or other clinical issues. Analysis of the clinical conditions of long-term patients also may identify those able to be released from hospitals. Cost analyses should compare the consumption of resources by outliers with those consumed by patients with more typical lengths of stay.

Analyzing the "tail" of the parameter estimates (particularly length of stay) frequency distributions would be rewarding, since for instance patients with extended



stays can utilize a disproportionate amount of hospital resources. Accordingly, analyzing consumption of the services for a few patients who stayed more than the typical LOS might reveal disproportionate use of health care resources.

In summary, distributions of LOS and number of hospitalizations are often markedly skewed to the right, rendering traditional parametric measures (e.g., mean and standard deviations) less useful, requiring that graphic visualization of the frequency distributions and other statistics (median, mode and geometric mean) be used to determine the typical length of stay (central tendency). Most importantly, as healthcare reform continues at an unfaltering pace, pressure will increase to reduce inpatient stays and minimize the use of costly services. Hospital morbidity data must be characterized appropriately.

### 6.1.3 Diagnosis Agreement

Valid epidemiologic studies using claims data are indeed possible, given that diagnoses are well coded and procedures associated with diagnoses are also reasonably well captured [Iezzoni LI, 1994; Fisher ES, 1992]. Using RAMQ physician claims that contain both diagnoses and procedures would result in a more complete and detailed information, making inferences easier. Any general inference regarding claims data without such complete information is simply unsupported; sensitivity and positive predictive value would greatly suffer from incomplete data (Appendix 4). Including procedure codes or linking claims records from multiple encounters may enhance the sensitivity of claims data for clinical conditions.

Advances in statistical methods have allowed investigators to better assess the quality of data by using different measures of agreement; however, little attention has been paid to the applicability of such measures in epidemiology [Sackett DL, 1991]. Researchers frequently rely on incomplete or inappropriate indices of agreement; they too rarely exploit new approaches for diagnostic tests and raters. Nevertheless, in our study, indices of agreement were generated with a significant understanding of their appropriateness and limitations. A prototype was the agreement of UGIB diagnostic codes in our analyses in which sample characteristics markedly affect the percentage agreement measure when determining efficiency (overall consistency) of accurate classifications. In this case, few individuals in the RAMQ medical services database

were identified as having had an upper gastrointestinal hemorrhage. With almost 99% of the answers concentrated in the lower right-hand corner (TN) of  $2 \times 2$  contingency table (as shown in Table 4.2), the efficiency for the UGIB diagnosis was high (99%). However, the well-known kappa measure ( $\kappa$ ) adjusted for this rare condition by using the potential agreement beyond chance as a baseline, and then expressing the actual agreement as a proportion of the potential agreement. As shown in previous studies, for relatively rare diagnoses, the proportion of true negative cases overwhelms the other cells in the contingency table leading to an artificially high percentage of efficiency [Robinson JR, 1997].

On the whole, results of our analyses revealed important characteristics of the developed algorithm and verified the overall agreement between the RAMQ medical services database and the hospital discharge summary, MED-ECHO, for different clinical conditions as being discussed in the following sections.

### **6.1.3.1 COPD Hospitalizations**

Patients in the training dataset (COPD cohort) were identified using a combination of diagnostic codes, as well as disease severity based on the drug utilizations. By applying the developed algorithm on the generated cohort, hospitalization episodes were estimated and compared to those observed in the gold

standard. Therefore, results of the agreement analysis were not only restricted to COPD diagnostic codes, but applicable to the utility of the selection criteria.

As quantified by the kappa index, the chance-corrected overall agreement of the RAMQ medical claims database with the gold standard identifying episodes of hospitalizations for COPD was found to be almost perfect with high sensitivity and specificity. Moreover, the positive predictive value of the algorithm was found to be high with a relatively low false positive rate. The algorithm used for diagnosis and severity of COPD was therefore found to be a good screening tool for COPD in the RAMQ database, a useful source for quantifying the burden of COPD in the community.

The results of this study are in total agreement with the previously published literature. A number of validation studies have found the recording of medical data in administrative databases to be nearly complete for chronic illnesses [Jick H, 1992; Jick H, 1991], but to date only the Saskatchewan database has demonstrated the internal validity of the recording of diagnosis of COPD, with a proportion of agreement of 64–88%, however no  $\kappa$  values were reported [Rawson NS and Malcolm E, 1995b]. Also in seven general practices in The Netherlands, GP diagnosis of chronic diseases ranged from a high of true positive diagnoses of 96% in diabetes mellitus to a low of 58% in chronic non-specific lung disease [Schellevis FG, 1993].

### 6.1.3.2 AMI Hospitalizations

Hospitalizations for AMI were identified using ICD-9 diagnosis code '410' so that the reliability of this code on the RAMQ claims data could be measured. According to the results, diagnostic code of AMI on RAMQ claims was reasonably reliable by which hospitalized patients with AMI were accurately identified about 70% of time (positive predictive value). Compared to COPD hospitalizations, the overall agreement between the RAMQ and MED-ECHO databases was found to be substantial for screening AMI hospitalizations with relatively lower sensitivity but still high specificity.

The lower sensitivity and high specificity suggested that in this analysis the RAMQ medical claims database was fairly accurate but some patients with an AMI diagnosis might end up with a different ICD-9 code. A common reason is that a patient might be admitted to a hospital for chest pain and physician has used a diagnosis code other than myocardial infarction (e.g., angina pectoralis); however, based on the complete medical and laboratory examinations, a diagnosis for AMI has been confirmed and listed in the patient's discharge summary, MED-ECHO database. Generally, errors resulted when the physician listed the acute myocardial infarction incorrectly, when a myocardial infarction occurred in a previous admission, or when myocardial infarction was "ruled out" (if it was the admitting diagnosis).

In this study, patients with an AMI diagnosis who were hospitalized less than 3 days were excluded because “rule-out” myocardial infarction has been a source of miscoding in prior studies [Petersen LA, 1999; Schiff GD and Yaacoub AS, 1989; Iezzoni LI, 1988]. Indeed, our findings are comparable to the results of other validation studies of AMI coding accuracy in both Canadian and American administrative databases [Fisher ES, 1992; Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group, 1989; Schiff GD and Yaacoub AS, 1989; Mascioli SR, 1989; Iezzoni LI, 1988; Kennedy GT, 1984]. In these studies, the sensitivity of using ICD-9 code 410 to detect AMI has been consistently greater than 80%. The specificity has varied from a low of 65% to 99%. The positive predictive value has also varied as a result, but has been consistently greater than 50%. In these studies, abstracting of medical records has been suggested as the only effective way in eliminating the false positive cases.

### **6.1.3.3 UGIB Hospitalizations**

Diagnostic data for conditions such as UGIB were substantially less reliable. In screening UGIB hospitalizations, reliance on a single set of diagnoses was no more optimal than reliance on one procedure (upper gastrointestinal endoscopy). Physician claims for gastrointestinal endoscopy within the first 24 hours of admission helped the

ascertainment of hospitalizations with an upper gastrointestinal hemorrhage. This perfectly indicates that in some settings, procedure fields on the hospital claims may be relevant, either in their own right or because they imply certain diagnoses of interest. In general, the low positive predictive value for screening UGIB hospitalizations may reflect the fact that these conditions are often difficult to diagnose and occasionally are confused with other gastrointestinal conditions or complications [Lawthers AG, 2000]. This is consistence with international studies and that “more complex cases are more prone to coding discrepancy” [MacIntyre CR, 1997a].

As evidenced in Appendix 4, the “count” of patients with hospitalizations for upper gastrointestinal bleeding was highly dependent upon which codes were selected. For instance, an UGIB hospitalization sample may contain a procedural code for upper gastrointestinal endoscopy within the first 24 hours of admission to the hospital (model **M1**), or a diagnosis code between 531.xx and 534.xx or 578.xx (Model **M2**). In the validation data, there were 8,899 individuals meeting the criteria for model **M1**, whereas this number was 9,034 for **M2**. However, in our analyses, in addition to a broad or high-sensitivity ICD-9 algorithm, other ICD-9 algorithms maximizing specificity (narrow, high-specificity algorithm) were also created (model **M3** and **M4**).

The tradeoff between sensitivity and specificity for the broad and narrow algorithms was quite apparent. By selecting the broad algorithm, sensitivity could be

augmented but, at the same time, false positive patients were increased threefold (from 1701 to 5144). If a “cleaner” sample of patients with UGIB hospitalizations is desired, the narrow algorithm can be used. However, the sample will be approximately two-third of all patients with UGIB hospitalizations (65% in this sample) and may potentially contain a selection bias if this smaller group is not representative of the larger population. These choices, and their effects, must be considered prior to patient selection, according to the objective(s) of the study.

In decision on the selection of the proper algorithm in epidemiologic studies, one should be aware that the actual balance between the true positive and false positive rates often matters [Glasziou P and Hilden J, 1989]. Whenever these rates are weighted differentially, both the prevalence and the conditional error rates have to be taken into consideration to make a balance decision. If the ruling-in or ruling-out of the target condition is the primary intended use of a diagnostic test, conditional indicators such as sensitivity and specificity still have to be used [Glas AS, 2003]. For instance, if the *risk ratio* (or in case-control studies “odds ratio”) is to be measured, there would be no bias in the measurement while using a test with higher specificity [White E, 1986; Lawrence C and Greenwald P, 1977]. However, the bias in the apparent *risk difference* will decrease as sensitivity increases (*ibid*).

Consistent with prior studies, our findings confirmed that identification of patients with difficult and more complicated clinical conditions by using ICD-9 codes can be difficult and fraught with unknown effects [Lawthers AG, 2000]. Using the



most specific algorithm, the overall agreement between the two databases was found to be substantial for screening UGIB hospitalizations with low sensitivity but rather high specificity. In general, sensitivity tended to be higher for algorithms with a broader diagnostic definition for UGIB hospitalizations.

However, summary measures such as kappa may hide the tradeoffs between sensitivity and specificity. Comparing the frequency with which a particular condition was noted in RAMQ medical services database and the frequency with which this condition was identified using other statistics (such as likelihood ratios and Kullback-Leibler distance) provided additional criteria for selection of the best UGIB identification algorithm. Likelihood ratios are often used to judge the rule-in and rule-out potentials once the rating results are known. However, to understand how an average hospitalization case would turn out (the '*before-test potential*'), KL distance was measured for all models. The use of this statistics, as the '*before-test*' potentials of ruling in and ruling out UGIB hospitalizations, also strongly pointed at the most specific (narrow definition) algorithm for identification of these events in the RAMQ database.

## 6.2 Limitations

Using administrative databases to carry out epidemiologic and outcomes studies, however, poses certain limitations. The most important limitation of our study is that the clinical accuracy of the coding can not be directly inferred from our findings.

Like previous studies of this kind, our study was designed to examine the accuracy with which physicians adhered to coding diagnosis, rather than to assess the validity of the claims data as a source of information on patients' health status or medical treatment. Consequently the reported findings do not directly answer the two major clinical relevant questions of whether a specific clinical condition present *in the patient* was coded on the computerized claims database (sensitivity) and whether specific conditions that coded actually were present *in the patient* (positive predictive value) [Fisher ES, 1992].

The distinction between coding conventions and patient's clinical status may explain lower accuracy with which some clinical conditions are coded. A diagnosis may have been presented but not recorded on the hospital discharge summary database because the condition was judged not to contribute to the care provided to the patient or it could be considered as rule-out diagnosis. Such misclassification

could affect both the sensitivity and positive predictive value of the coded diagnoses leading to an underestimation of these measures.

In addition, our study represents a cross-sectional comparison of RAMQ physician claims and hospital discharge summary diagnosis data because we only examined the first matching record for each patient. In many claims datasets, information may exist from a number of health care encounters for a given patient. By combining claims information from more than one encounter, a longitudinal claims record can be constructed that may better identify clinical conditions (case ascertainment). For patients who had multiple hospital encounters, researchers would have additional information to specify illness, leading to a higher level of accuracy.

Furthermore, other limitation is caused by the fact that RAMQ physician claims database rarely includes all information of specific interest for epidemiologic studies and outcomes research (e.g., pharmaceutical services). However, linking two or more RAMQ databases would decrease this limitation allowing investigators to answer research objectives with an even greater scope, volume, and depth. The usefulness of such linkage was clearly shown in the training dataset where COPD patients were selected on the basis of all available information including pharmaceutical services provided.

Moreover, it should be kept in mind that in Canada, billing claims for physician services typically provide complete capture of procedure codes; whereas,

most provinces do not require a diagnosis on claims for payment, so it is essentially an optional field on the claim [Williams JI and Young W, 1996a]. Nevertheless, these codes may not necessarily match those used in hospital records, leading to a lower level of concordance and greater discrepancy between RAMQ and MED-ECHO databases.

Finally, the validation study on the RAMQ database has been performed using elderly patient population (age  $\geq 65$ ) and the results might not be applicable to younger populations, particularly children. Further studies are required to establish the validity of the RAMQ medical claims database to identify episodes of hospitalizations in younger age groups.

## 6.3 Implications

Results of the analyses performed in this study established the RAMQ medical services database as being an excellent source of information with a considerable potential for the study of the utilization of health care resources (particularly hospitalizations) and its outcomes more efficiently and in a shorter time.

The potential advantages of using insurance claims datasets, including RAMQ databases, for clinical research have been described in many previous publications [Lohr KN, 2000; Fisher ES, 1992; Lohr KN, 1990]. They include large samples of geographically dispersed patients; longitudinal records; data already collected and available; and defined sampling frames (*ibid*).

In addition, accessibility to the RAMQ databases and continuity of information are of great interest. All information is readily available for researchers; RAMQ databases are easily accessible at relatively low cost and within reasonable time, compared to other data sources (including MED-ECHO discharge summaries). Continuity of data rarely receives enough attention. Although it is not a priority issue, but seems to be challenging for some researchers while they conduct studies for which the latest (up-to-date) information is required. This problem can be solved by using RAMQ databases, since claims are constantly recorded in these databases, not at the end of each fiscal year (as seen in the case of MED-ECHO).

More importantly, the possibility of direct linkage between physician claims and drugs prescribed offers the unique opportunity to examine clinical practice pattern, as well as the patterns of health care resources and drug consumptions. In our analyses, it was shown that multiple claims data streams would result in a better case ascertainment. For instance, unlike AMI and UGIB cases, COPD hospitalizations were identified not only based on the presence of claimed diagnoses on patients' records, but also on a combination of certain inclusion and exclusion criteria, as well as the history of drug utilization available in pharmaceutical services database (through record linkage).

Overall, in a single-payer environment such as Quebec health care system, administrative data provide a powerful resource for population-based evaluation of the burden of different diseases. Although hospital discharge summary data are considered the gold standard (in the absence of chart review) for the analyses of hospital utilization, variation, and trend over time; its timeliness is reduced by delays in submissions before the database is complete for a given year. The current study demonstrated that RAMQ databases could be used to measure the episodes of hospitalizations for different diseases; it could provide reasonable diagnostic agreement with data collected in the hospital discharge summary. In addition, RAMQ data provide an efficient tool for measuring hospitalization trends. The time and cost of future studies can drastically be reduced by using the RAMQ health administrative

databases as the source of data; access to hospitalization records is no longer required through a linkage to MED-ECHO.

Although the algorithm with a gap of 7 days was found to be the best algorithm identifying the episodes of hospitalizations in the RAMQ, for patients with the clinical conditions that require multiple hospitalizations or longer duration of stay, this algorithm might result in better estimates.

# ***Conclusion***

---

# **7**

While estimates of in-hospital episodes relying extensively on discharge abstracts would provide more accurate information, the associated time for accessing this information currently appear prohibitive in many studies. Time-efficiency in using RAMQ claims data provided the key motivation for this dissertation, which explored the capacity of claims data, a less costly, and albeit less informative data source, to estimate hospital morbidities, including LOS and readmissions.

In summary, using MED-ECHO database as the gold standard, as the specificity of claims data increased; the likelihood that the claims data would overestimate the number of patients who satisfy the eligible or passing criteria decreases. Existing evidence also suggested that false positive diagnostic codes in the RAMQ physician claims were rare (high specificity), which is consistent with previous studies [Robinson JR, 1997; Jollis JG, 1993]. As documented, when false positive diagnoses presented they tended to be attributed to random errors in coding or artifacts created by the use of different time windows of measurement, such as dates of admission and discharge (*ibid*), as was believed in our analyses.

The use of RAMQ medical services database lends itself readily to analyses, providing extremely large data sets, and at very little cost. However, validation of



RAMQ database regarding episodes of hospitalizations was essential since access to the MED-ECHO files and collection of resource data is often time consuming and implies inefficiency. Substantial time and cost savings can be made if validated information could instead be gathered from RAMQ. Data provided by the RAMQ databases contain a wealth of information on resource utilization and outcome that are used for reimbursement, quality monitoring, and research.

Overall, this study described the presence and magnitude of data strengths and weaknesses in the RAMQ medical services database in comparison to the MED-ECHO data, allowing a better interpretation of the future studies from the RAMQ databases. The disagreements in the diagnoses identifying in-hospital episodes continue to be a concern and should be the focus of efforts to further improve RAMQ database.

The developed algorithm would provide a feasible and convenient approach to identify hospitalization episodes in future epidemiologic, economic, and outcomes research. Although it was initially developed for a specific purpose (COPD), it may also be used in studies of other clinical events, as confirmed in the analyses of AMI and UGIB hospitalizations. With some modifications of the codes for a diagnosis of interest, it may be used to study trends of hospitalizations or the relation between patients' characteristics and variation in LOS in other diseases.

In conclusion, the developed algorithm is a valuable tool for identification and evaluation of inhospital episodes in the RAMQ medical services database. It can easily be integrated in epidemiologic and outcomes research studies.

As suggested by the results of this study, physicians and other health professionals must be fastidious in recording relevant information, and be educated about some of the controversies in disease and procedure classification. The quality of coding for diagnoses and procedures must be improved to the greatest extent possible. Continuing education and regular audits, which are already integral to the operations of health records departments of large hospitals, must be augmented and extended to institutions of all sizes.

A substantial amount of claims data research has been done. In the future, the assessment of the quality of care will become increasingly important due to the increasing need for a better understanding of the process and results of health care. The increasing role of claims data for research can be supported by the systematic analyses of the appropriateness of these databases for different research questions. In addition, data validation methods to assess the quality of available information, particularly administrative databases, are still a field needing much further research.

# Reference

---

**American Thoracic Society.** *Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease.* American Thoracic Society. Am.J.Respir.Crit Care Med. 1995; **152** RF(5 Pt 2): S77-121.

**Anderson GF.** *Monitoring hospital readmissions.* Cleve.Clin.J.Med. 1989; **56**(7): 674-675.

**Anderson GF and Steinberg EP.** *Hospital readmissions in the Medicare population.* N.Engl.J.Med. 1984; **311**(21): 1349-1353.

**Anderson GF and Steinberg EP.** *Predicting hospital readmissions in the Medicare population.* Inquiry 1985; **22**(3): 251-258.

**Angus DC, Linde-Zwirble WT, Sirio CA, Rotondi AJ, Chelluri L, Newbold RC, III, Lave JR, Pinsky MR.** *The effect of managed care on ICU length of stay: implications for medicare.* JAMA 1996; **276**(13): 1075-1082.

**Ashton CM, Del Junco DJ, Soucek J, Wray NP, Mansyur CL.** *The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence.* Med.Care 1997; **35**(10): 1044-1059.

**Ashton CM, Kuykendall DH, Johnson ML, Wray NP, Wu L.** *The association between the quality of inpatient care and early readmission.* Ann.Intern.Med. 1995; **122**(6): 415-421.

**Attena F, Agozzino E, Troisi MR, Granito C, Del Prete U.** *Appropriateness of admission and hospitalization days in a specialist hospital.* Ann.Ig 2001; **13**(2): 121-127.

**Austin PC, Daly PA, Tu JV.** *A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario.* Am Heart J 2002; **144**(2): 290-296.

**Bare ML, Prat A, Lledo L, Asenjo MA, Salleras L.** *Appropriateness of admissions and hospitalization days in an acute-care teaching hospital.* Rev.Epidemiol.Sante Publique 1995; **43**(4): 328-336.

**Beggs VL, Birkemeyer NJ, Nugent WC, Dacey LJ, O'Connor GT.** *Factors related to rehospitalization within thirty days of discharge after coronary artery bypass grafting.* Best.Pract.Benchmarking.Healthc. 1996; **1**(4): 180-186.

**Beland F.** *[Systematic bias in measuring of the utilization of ambulatory medical services].* Can.J Public Health 1989: 58-62.

**Berki SE, Ashcraft ML, Newbrander WC.** *Length-of-stay variations within ICDA-8 diagnosis-related groups.* Med.Care 1984; **22**(2): 126-142.

**Bernstein SJ and Hilborne LH.** *Clinical indicators: the road to quality care?* Jt.Comm J Qual.Improv. 1993; **19**(11): 501-509.

**Berry CC.** *The kappa statistic.* JAMA 1992; **268**(18): 2513-2514.

**Berwick DM.** *Health services research and quality of care. Assignments for the 1990s.* Med.Care 1989; **27**(8): 763-771.

**Berwick DM and Wald DL.** *Hospital leaders' opinions of the HCFA mortality data.* JAMA 1990; **263**(2): 247-249.

**Bigby J, Dunn J, Goldman L, Adams JB, Jen P, Landefeld CS, Komaroff AL.** *Assessing the preventability of emergency hospital admissions. A method for evaluating the quality of medical care in a primary care facility.* Am.J Med. 1987; **83**(6): 1031-1036.

**Birnbaum HG, Cremieux PY, Greenberg PE, LeLorier J, Ostrander JA, Venditti L.** *Using healthcare claims data for outcomes research and pharmacoeconomic analyses.* Pharmacoeconomics. 1999; **16**(1): 1-8.

**Blais L, Boucher JM, Couture J, Rahme E, LeLorier J.** *Impact of a cost-sharing drug insurance plan on drug utilization among older people.* J.Am.Geriatr.Soc. 2001; **49**(4): 410-414.

**Blais L, Bourbeau J, Sheehy O, LeLorier J.** *Inhaled Corticosteroids in COPD: Determinants of Use and Trends in Patient Persistence on Treatment.* Can Respir J. 2004; **11**(1): 27-32.

**Blais L, Couture J, Rahme E, LeLorier J.** *Impact of a cost sharing drug insurance plan on drug utilization among individuals receiving social assistance.* Health Policy 2003; **64**(2): 163-172.

**Blais R, Breton JJ, Fournier M, St Georges M, Berthiaume C.** *Are mental health services for children distributed according to needs?* Can J Psychiatry 2003; **48**(3): 176-186.

**Blais R and Maiga A.** *Do ethnic groups use health services like the majority of the population? A study from Quebec, Canada.* Soc.Sci.Med. 1999; **48**(9): 1237-1245.

**Brackett MH.** *The data warehouse challenge: Taming data chaos.* 1996.

**Briggs DD, Kuritzky L, Boland C.** *Early detection and management of COPD, recommendations from the National COPD Awareness Panel.* J Resp Dis 2000; **21**(suppl): S1-S17.

**Bright RA, Avorn J, Everitt DE.** *Medicaid data as a resource for epidemiologic studies: strengths and limitations.* J Clin.Epidemiol. 1989; **42**(10): 937-945.

**Brook RH and Appel FA.** *Quality-of-care assessment: choosing a method for peer review.* N.Engl.J Med. 1973; **288**(25): 1323-1329.

**Brook RH, McGlynn EA, Cleary PD.** *Quality of health care. Part 2: measuring quality of care.* N.Engl.J Med. 1996; **335**(13): 966-970.

**Byrt T, Bishop J, Carlin JB.** *Bias, prevalence and kappa.* J Clin.Epidemiol. 1993; **46**(5): 423-429.

**Calle JE, Saturno PJ, Parra P, Rodenas J, Perez MJ, Eustaquio FS, Aguinaga E.** *Quality of the information contained in the minimum basic data set: results from an evaluation in eight hospitals.* Eur.J Epidemiol. 2000; **16**(11): 1073-1080.

**Canadian Institute for Health Information.** *Health Care in Canada.* 2002.

**Canadian Institute for Health Information (CIHI).** *Comparative Analysis Report: Hospital Utilization in Canada.* Ottawa, ON: Canadian Institute for Health Information. 1994.

**Canadian Respiratory Review Panel.** *Guidelines for the Treatment of Chronic Obstructive Pulmonary Disease (COPD).* 1998.

**Caplan RA, Posner KL, Cheney FW.** *Effect of outcome on physician judgments of appropriateness of care.* JAMA 1991; **265**(15): 1957-1960.

**Chen E.** *Reliability of HMRI data among selected obstetrical population.* ICES, North York. 1995.

**Choinière R, Pageau M, Ferland M.** *Prevalence and Geographic Disparities in Certain Congenital Anomalies in Quebec: Comparison of Estimation Methods.* Chronic.Dis.Can 1999; **20**(2): 51-57.

**Chow SC and Tse SK.** *Outlier detection in bioavailability/bioequivalence studies.* Stat Med 1990; **9**(5): 549-558.

**Chowdhury S, Bodemar G, Haug P, Babic A, Wigertz O.** *Methods for knowledge extraction from a clinical database on liver diseases.* Comput.Biomed.Res. 1991; **24**(6): 530-548.

**Cicchetti DV and Feinstein AR.** *High agreement but low kappa: II. Resolving the paradoxes.* J.Clin.Epidemiol. 1990; **43**(6): 551-558.

**Cleary PD, Greenfield S, Mulley AG, Pauker SG, Schroeder SA, Wexler L, McNeil BJ.** *Variations in length of stay and outcomes for six medical and surgical conditions in Massachusetts and California.* JAMA 1991; **266**: 73-79.

**Clewer A and Perkins D.** *Economics for health care management.* 1998.

**Cohen MM.** *Using administrative data for case-control studies: the case of the Papanicolaou smear.* Ann Epidemiol. 1993; **3**(1): 93-98.

**Connell FA, Diehr P, Hart LG.** *The use of large data bases in health care studies.* Annu.Rev.Public Health 1987; **8**: 51-74.

**Cooper GS, Chak A, Lloyd LE, Yurchick PJ, Harper DL, Rosenthal GE.** *The accuracy of diagnosis and procedural codes for patients with upper GI hemorrhage.* Gastrointest.Endosc. 2000; **51**(4 Pt 1): 423-426.

**Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA.** *The sensitivity of Medicare claims data for case ascertainment of six common cancers.*

**Corrigan JM and Martin JB.** *Identification of factors associated with hospital readmission and development of a predictive model.* Health Serv.Res. 1992; **27**(1): 81-101.

**Couris CM, Colin C, Rabilloud M, Schott AM, Ecochard R.** *Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases.* J Clin.Epidemiol. 2002; **55**(4): 386-391.

**Covey MK, Larson JL, Alex CG, Wirtz S, Langbein WE.** *Test-retest reliability of symptom-limited cycle ergometer tests in patients with chronic obstructive pulmonary disease.* Nurs.Res. 1999; **48**(1): 9-19.

**Dandona L, Morrison JC, Robin AL, Quigley HA.** *"Outlier" intraocular pressure readings: valid or invalid for statistical analysis?* Ophthalmic Surg. 1989; **20**(10): 737-743.

**Davidson W, Molloy DW, Somers G, Bedard M.** *Relation between physician characteristics and prescribing for elderly people in New Brunswick.* CMAJ. 1994; **150**(6): 917-921.

**Delfino RJ, Becklake MR, Hanley JA.** *Reliability of hospital data for population-based studies of air pollution.* Arch.Environ.Health 1993; **48**(3): 140-146.

**Demers M.** *Frequent users of ambulatory health care in Quebec: the case of doctor-shoppers.* CMAJ. 1995; **153**(1): 37-42.

**Demlo LK and Campbell PM.** *Improving hospital discharge data: lessons from the National Hospital Discharge Survey.* Med.Care 1981; **19**(10): 1030-1040.

**Demlo LK, Campbell PM, Brown SS.** *Reliability of information abstracted from patients' medical records.* Med Care 1978; **16**(12): 995-1005.

**Deyo RA, Cherkin DC, Ciol MA.** *Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases.* J.Clin.Epidemiol. 1992; **45**(6): 613-619.

**Dickerson RN, Gervasio JM, Riley ML, Murrell JE, Hickerson WL, Kudsk KA, Brown RO.** *Accuracy of predictive methods to estimate resting energy expenditure of thermally-injured patients.* JPEN J Parenter. Enteral Nutr. 2002; **26**(1): 17-29.

**Domingo-Salvany A, Lamarca R, Ferrer M, Garcia-Aymerich J, Alonso J, Felez M, Khalaf A, Marrades RM, Monso E, Serra-Batlles J, Anto JM.** *Health-related quality of life and mortality in male patients with chronic obstructive pulmonary disease.* Am.J Respir.Crit Care Med. 2002; **166**(5): 680-685.

**Donabedian A.** *The quality of care. How can it be assessed?* JAMA 1988; **260**(12): 1743-1748.

**Donabedian A.** *The end results of health care: Ernest Codman's contribution to quality assessment and beyond.* Milbank Q. 1989; **67**(2): 233-256.

**Dubois RW and Brook RH.** *Preventable deaths: who, how often, and why?* Ann.Intern.Med. 1988; **109**(7): 582-589.

**Dubois RW, Brook RH, Rogers WH.** *Adjusted hospital death rates: a potential screen for quality of medical care.* Am.J Public Health 1987a; **77**(9): 1162-1166.

**Dubois RW, Rogers WH, Moxley JH, III, Draper D, Brook RH.** *Hospital inpatient mortality. Is it a predictor of quality?* N.Engl.J Med. 1987b; **317**(26): 1674-1680.

**Dussaucy A, Viel JF, Mulin B, Euvrard J.** *[The framework Prospective Payment Information Systems: bias, sources of errors and consequences].* Rev.Epidemiol.Sante Publique 1994; **42**(4): 345-358.

**Edouard L and Rawson NS.** *Reliability of the recording of hysterectomy in the Saskatchewan health care system.* Br.J.Obstet.Gynaecol. 1996; **103**(9): 891-897.



**Edwards WH, Sr., Edwards WH, Jr., Martin RS, III, Mulherin JL, Jr., Bullock D.** *Resource utilization and pathways: meeting the challenge of cost containment.* Am.Surg. 1996; **62**(10): 830-834.

**Emanuel EJ and Emanuel LL.** *What is accountability in health care?* Ann.Intern.Med. 1996; **124**(2): 229-239.

**Engels JM and Diehr P.** *Imputation of missing longitudinal data: a comparison of methods.* J Clin.Epidemiol. 2003; **56**(10): 968-976.

**Epstein AM, Stern RS, Weissman JS.** *Do the poor cost more? A multihospital study of patients' socioeconomic status and use of hospital resources.* N.Engl.J.Med. 1990; **322**(16): 1122-1128.

**Esmail A, Quayle JA, Roberts C.** *Assessing the appropriateness of paediatric hospital admissions in the United Kingdom.* J Public Health Med. 2000; **22**(2): 231-238.

**Ettner SL and Notman EH.** *How well do ambulatory care groups predict expenditures on mental health and substance abuse patients?* Adm Policy Ment.Health 1997; **24**(4): 339-357.

**Evans RL, Hendricks RD, Lawrence KV, Bishop DS.** *Identifying factors associated with health care use: a hospital-based risk screening index.* Soc.Sci.Med. 1988; **27**: 947-954.

**Fan VS, Curtis JR, Tu SP, McDonell MB, Fihn SD.** *Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases.* Chest 2002; **122**(2): 429-436.

**Feinstein AR.** *ICD, POR, and DRG. Unsolved scientific problems in the nosology of clinical medicine.* Arch.Intern.Med. 1988; **148**(10): 2269-2274.

**Feinstein AR and Cicchetti DV.** *High agreement but low kappa: I. The problems of two paradoxes.* J.Clin.Epidemiol. 1990; **43**(6): 543-549.

**Ferguson GT and Cherniack RM.** *Management of chronic obstructive pulmonary disease.* N.Engl.J.Med. 1993; **328**(14): 1017-1022.

**Fethke CC, Smith IM, Johnson N.** *"Risk" factors affecting readmission of the elderly into the health care system.* Med.Care 1986; **24**(5): 429-437.

**Finlayson G, Watson D, Jacobs P.** *The relative cost of inpatient care in Manitoba's largest hospitals.* Healthc.Manage.Forum 2002; **Suppl**: 47-52.

**Fisher ES, Wennberg JE, Stukel TA, Sharp SM.** *Hospital readmission rates for cohorts of Medicare beneficiaries in Boston and New Haven.* N.Engl.J Med. 1994; **331**(15): 989-995.

**Fisher ES, Whaley FS, Krushat WM, Malenka DJ, Fleming C, Baron JA, Hsia DC.** *The accuracy of Medicare's hospital claims data: progress has been made, but problems remain.* Am.J.Public Health 1992; **82**(2): 243-248.

**Fleming C, Fisher ES, Chang CH, Bubolz TA, Malenka DJ.** *Studying outcomes and hospital utilization in the elderly. The advantages of a merged data base for Medicare and Veterans Affairs hospitals.* Med.Care 1992; **30**(5): 377-391.

**Fleming ST, Hicks LL, Bailey RC.** *Interpreting the Health Care Financing Administration's mortality statistics.* Med.Care 1995; **33**(2): 186-201.

**Fowles JB, Fowler E, Craft C, McCoy CE.** *Comparing claims data and self-reported data with the medical record for Pap smear rates.* Eval.Health Prof. 1997; **20**(3): 324-342.

**Garbe E, Boivin JF, LeLorier J, Suissa S.** *Systematic bias in measuring of the utilization of ambulatory medical services.* Can J Public Health 1998; **80**(1): 58-62.

**Garbe E, LeLorier J, Boivin JF, Suissa S.** *Risk of ocular hypertension or open-angle glaucoma in elderly patients on or. Lancet 1997; **350**(9083): 979-82.*

**Garnick DW, DeLong ER, Luft HS.** *Measuring hospital mortality rates: are 30-day data enough? Ischemic Heart Disease Patient Outcomes Research Team.* Health Serv.Res. 1995; **29**(6): 679-695.

**Gerald LB, Sanderson B, Redden D, Bailey WC.** *Chronic obstructive pulmonary disease stage and 6-minute walk outcome.* J.Cardiopulm.Rehabil. 2001; **21**(5): 296-299.

**Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM.** *The diagnostic odds ratio: a single indicator of test performance.* J Clin.Epidemiol. 2003; **56**(11): 1129-1135.

**Glasziou P and Hilden J.** *Test selection measures.* Med Decis.Making 1989; **9**(2): 133-141.

**Gold L, Parry D, Raftery J, Stevens A, Walshe K, Woolley M.** *An assessment of potential indicators of clinical effectiveness.* 1998.

**Goldfarb MG, Hornbrook MC, Higgins CS.** *Determinants of hospital use: a cross-diagnostic analysis.* Med.Care 1983; **21**(1): 48-66.

**Green J, Passman LJ, Wintfeld N.** *Analyzing hospital mortality. The consequences of diversity in patient mix.* JAMA 1991; **265**(14): 1849-1853.

**Green J and Wintfeld N.** *How accurate are hospital discharge data for evaluating effectiveness of care?* Med.Care 1993(31): 719-731.

**Greenfield S.** *The state of outcome research: are we on target?* N.Engl.J.Med. 1989; **320**(17): 1142-1143.

**Greenfield S, Apolone G, McNeil BJ, Cleary PD.** *The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. Comorbidity and outcomes after hip replacement.* Med.Care 1993; **31**(2): 141-154.

**Greenland S and Robins JM.** *Confounding and misclassification.* Am J Epidemiol. 1985; **122**(3): 495-506.

**Halfon P and Eggli Y.** *Screening inappropriate hospital days on the basis of routinely available data.* Int.J Qual.Health Care 2001; **13**(4): 289-299.

**Hannan EL, Kilburn H, Jr., Lindsey ML, Lewis R.** *Clinical versus administrative data bases for CABG surgery. Does it matter?* Med Care 1992; **30**(10): 892-907.

**Hannan EL, Racz MJ, Jollis JG, Peterson ED.** *Using Medicare claims data to assess provider quality for CABG surgery: does it work well enough?* Health Serv.Res. 1997; **31**(6): 659-678.

**Hannan EL, Siu AL, Kumar D, Kilburn H, Jr., Chassin MR.** *The decline in coronary artery bypass graft surgery mortality in New York State. The role of surgeon volume.* JAMA 1995; 273(3): 209-213.

**Harris HE, Ellison GT, Holliday M, Lucassen E.** *Methodological considerations in the design of an obstetric database abstracted from medical records.* Methods Inf.Med. 1997; 36(3): 191-200.

**Harrison ML, Graff LA, Roos NP, Brownell MD.** *Discharging patients earlier from Winnipeg hospitals: does it adversely affect quality of care?* CMAJ. 1995; 153(6): 745-751.

**Hartz AJ, Kuhn EM, Yuan Z, Bailey RC, Rimm AA.** *Adequacy of national mortality data for internal hospital use: a case study.* Int.J Qual.Health Care 1995; 7(2): 109-118.

**Hawker GA, Coyte PC, Wright JG, Paul JE, Bombardier C.** *Accuracy of administrative data for assessing outcomes after knee replacement surgery.* J.Clin.Epidemiol. 1997; 50(3): 265-273.

**Hayward RA, Bernard AM, Rosevear JS, Anderson JE, McMahon LF, Jr.** *An evaluation of generic screens for poor quality of hospital care on a general medicine service.* Med.Care 1993; 31(5): 394-402.

**Health Canada.** *Economic Burden of Illness in Canada.* 1998.

**Health Canada.** *Editorial Board Respiratory Disease in Canada.* 2001.

**Henderson J, Goldacre MJ, Graveney MJ, Simmons HM.** *Use of medical record linkage to study readmission rates.* BMJ 1989a; 299(6701): 709-713.

**Henderson J, Goldacre MJ, Griffith M, Simmons HM.** *Day case surgery: geographical variation, trends and readmission rates.* J Epidemiol.Community Health 1989b; 43(3): 301-305.

**Hogg RV and Tanis EA.** *Probability and statistical inference.* 1988.

**Horner RD, Paris JA, Purvis JR, Lawler FH.** *Accuracy of patient encounter and billing information in ambulatory care.* J Fam.Pract. 1991; 33(6): 593-598.

**Hsia DC, Ahern CA, Ritchie BP, Moscoe LM, Krushat WM.** *Medicare reimbursement accuracy under the prospective payment system, 1985 to 1988.* JAMA 1992; **268**(7): 896-899.

**Huff ED.** *Utility and limitations of claims data.* Ann.Intern.Med. 1994; **120**(12): 1049-1.

**Ibrahim JE, Majoor JW, Cicuttinni F, McNeil JJ.** *Readmission as an outcome variable.* Med.J Aust. 1999; **171**(4): 218-219.

**Iezzoni LI.** *Severity of illness measures. Comments and caveats.* Med.Care 1990a; **28**(9): 757-761.

**Iezzoni LI.** *Using administrative diagnostic data to assess the quality of hospital care. Pitfalls and potential of ICD-9-CM.* Int.J.Technol.Assess.Health Care 1990b; **6**(2): 272-281.

**Iezzoni LI.** *Risk Adjustment for Medical Outcomes Study.* In: Medical Effectiveness Research.Data and Methods 1992: 83-97.

**Iezzoni LI.** *Assessing quality using administrative data.* Ann.Intern.Med. 1997a; **127**(8 Pt 2): 666-674.

**Iezzoni LI, Ash AS, Shwartz M, Mackiernan YD.** *Differences in procedure use, in-hospital mortality, and illness severity by gender for acute myocardial infarction patients: are answers affected by data source and severity measure?* Med.Care 1997b; **35**(2): 158-171.

**Iezzoni LI, Burnside S, Sickles L, Moskowitz MA, Sawitz E, Levine PA.** *Coding of acute myocardial infarction. Clinical and policy implications.* Annals of Internal Medicine 1988; **109**(9): 745-751.

**Iezzoni LI, Daley J, Heeren T, Foley SM, Fisher ES, Duncan C, Hughes JS, Coffman GA.** *Identifying complications of care using administrative data.* Med.Care 1994; **32**(7): 700-715.

**Iron K, Goel V, Williams JI.** *Concordance of hospital discharge abstracts and physician claims for surgical procedures in Ontario.* ICES, North York. 1995.

**Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J.** *Interpreting hospital mortality data. The role of clinical risk adjustment.* JAMA 1988a; **260**(24): 3611-3616.

**Jencks SF, Williams DK, Kay TL.** *Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities.* JAMA 1988b; **260**(15): 2240-2246.

**Jick H, Jick SS, Derby LE.** *Validation of information recorded on general practitioner based computerised data resource in the United Kingdom.* BMJ 1991; **302**(6779): 766-768.

**Jick H, Terris BZ, Derby LE, Jick SS.** *Further validation of information recorded on a general practitioner based computerised data resource in the United Kingdom.* Pharmacoepidemiol Drug Safety 1992; **1**: 347-349.

**Johnson TS, Engstrom JL, Haney SL, Mulcrone SL.** *Reliability of three length measurement techniques in term infants.* Pediatr.Nurs. 1999; **25**(1): 13-17.

**Jollis JG and Romano PS.** *Pennsylvania's Focus on Heart Attack--grading the scorecard.* N.Engl.J.Med 1998; **338**(14): 983-987.

**Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB.** *Discordance of Databases Designed for Claims Payment versus Clinical Information Systems: Implications for Outcomes Research.* Annals of Internal Medicine 1993; **119**(8): 844-850.

**Kahn KL, Brook RH, Draper D, Keeler EB, Rubenstein LV, Rogers WH, Kosecoff J.** *Interpreting hospital mortality data. How can we proceed?* JAMA 1988; **260**(24): 3625-3628.

**Kashner TM.** *Agreement between administrative files and written medical records: a case of the Department of Veterans Affairs.* Med Care 1998; **36**(9): 1324-1336.

**Kassirer JP.** *The quality of care and the quality of measuring it.* N.Engl.J.Med. 1993; **329**(17): 1263-1265.

**Katz JN, Wright EA, Guadagnoli E, Liang MH, Karlson EW, Cleary PD.** *Differences between men and women undergoing major orthopedic surgery for degenerative arthritis.* Arthritis Rheum. 1994; 37(5): 687-694.

**Kennedy GT, Stern MP, Crawford MH.** *Miscoding of hospital discharges as acute myocardial infarction: implications for surveillance programs aimed at elucidating trends in coronary artery disease.* Am J Cardiol. 1984; 53(8): 1000-1002.

**Kephart G, Sketris I, Smith M, Maheu A, Brown M.** *Coprescribing of nonsteroidal anti-inflammatory drugs and cytoprotective and antiulcer drugs in Nova Scotia's senior population.* Clin.Ther. 1995; 17(6): 1159-1173.

**Knaus WA, Wagner DP, Zimmerman JE, Draper EA.** *Variations in mortality and length of stay in intensive care units.* Ann.Intern.Med. 1993; 118(10): 753-761.

**Kozyrskyj AL and Mustard CA.** *Validation of an electronic, population-based prescription database.* The Annals of Pharmacotherapy 1998; 32(11): 1152-1157.

**Kraemer HC.** *Evaluating Medical Tests.* 1992.

**Kraemer HC and Bloch DA.** *A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation.* Stat.Med. 1994; 13(8): 876-880.

**Lacasse Y, Brooks D, Goldstein RS.** *Trends in the epidemiology of COPD in Canada, 1980 to 1995. COPD and Rehabilitation Committee of the Canadian Thoracic Society.* Chest 1999; 116(2): 306-313.

**Lahad A and Yodfat Y.** *Impact of comorbidity on well-being in hypertension: case control study.* J.Hum.Hypertens. 1993; 7(6): 611-614.

**Landis JR and Koch GG.** *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers.* Biometrics 1977a; 33(2): 363-374.

**Landis JR and Koch GG.** *The measurement of observer agreement for categorical data.* Biometrics 1977b; 33(1): 159-174.

**Lang T, Davido A, Logerot H, Meyer L.** *Appropriateness of admissions: the French experience.* Int.J Qual.Health Care 1995; 7(3): 233-238.

**Langford AM and Reitz JA.** *Hospital use of clinical and organizational performance indicators.* 1995(8).

**Lawrence C and Greenwald P.** *Epidemiologic screening: a method to add efficiency to epidemiologic research.* Am J Epidemiol. 1977; 105(6): 575-581.

**Lawthers AG, McCarthy EP, Davis RB, Peterson LE, Palmer RH, Iezzoni LI.** *Identification of in-hospital complications from claims data. Is it valid?* Med Care 2000; 38(8): 785-795.

**Leape LL, Hilborne LH, Park RE, Bernstein SJ, Kamberg CJ, Sherwood M, Brook RH.** *The appropriateness of use of coronary artery bypass graft surgery in New York State.* JAMA 1993; 269(6): 753-760.

**Lee WC.** *Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance.* Int.J Epidemiol. 1999; 28(3): 521-525.

**LeLorier J.** *Patterns of prescription of nonsteroidal antiinflammatory drugs and gastroprotective agents.* J Rheumatol.Suppl 1995; 43: 26-27.

**Levy AR, Mayo NE, Grimard G.** *Rates of transcervical and pertrochanteric hip fractures in the province of Quebec, Canada, 1981-1992.* Am.J Epidemiol. 1995; 142(4): 428-436.

**Levy AR, Tamblyn RM, Fitchett D, Mcleod PJ, Hanley JA.** *Coding accuracy of hospital discharge data for elderly survivors of myocardial infarction.* Can J Cardiol. 1999; 15(11): 1277-1282.

**Lewis NJ, Patwell JT, Briesacher BA.** *The role of insurance claims databases in drug therapy outcomes research.* Pharmacoeconomics. 1993; 4(5): 323-330.

**Librero J, Peiro S, Ordinana R.** *Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days.* J.Clin.Epidemiol. 1999; 52(3): 171-179.

**Linn BS, Linn MW, Gurel L.** *Cumulative illness rating scale.* J.Am.Geriatr.Soc. 1968; 16(5): 622-626.



**Lloyd SS and Rissing JP.** *Physician and coding errors in patient records.* JAMA 1985; **254**(10): 1330-1336.

**Lohr KN.** *Use of insurance claims data in measuring quality of care.* Int.J Technol.Assess.Health Care 1990; **6**(2): 263-271.

**Lohr KN.** *Health outcomes methodology symposium: summary and recommendations.* Med.Care 2000; **38**(9 Suppl): II194-II208.

**Lohr KN, Donaldson MS, Harris-Wehling J.** *Medicare: a strategy for quality assurance, V: Quality of care in a changing health care environment.* QRB Qual.Rev.Bull. 1992; **18**(4): 120-126.

**Lombrail P, Minvielle E, Comar L, Gottot S.** *[Prospective Payment Information Systems and epidemiology: a difficult link to establish].* Rev.Epidemiol.Sante Publique 1994; **42**(4): 334-344.

**Loop FD, Christiansen EK, Lester JL, Cosgrove DM, Franco I, Golding LR.** *A strategy for cost containment in coronary surgery.* JAMA 1983; **250**(1): 63-66.

**Ludke RL, MacDowell NM, Booth BM, Hunter SA.** *Appropriateness of admissions and discharges among readmitted patients.* Health Serv.Res. 1990; **25**(3): 501-525.

**MacIntyre CR, Ackland MJ, Chandraraj EJ, Pilla JE.** *Accuracy of ICD-9-CM codes in hospital morbidity data, Victoria: implications for public health research.* Aust.N.Z.J Public Health 1997a; **21**(5): 477-482.

**MacIntyre CR, Brook CW, Chandraraj E, Plant AJ.** *Changes in bed resources and admission patterns in acute public hospitals in Victoria, 1987-1995.* Med.J Aust. 1997b; **167**(4): 186-189.

**Mackinnon A.** *A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement.* Comput.Biol.Med 2000; **30**(3): 127-134.

**Malenka DJ, McLerran D, Roos N, Fisher ES, Wennberg JE.** *Using administrative data to describe casemix: a comparison with the medical record.* J.Clin.Epidemiol. 1994; **47**(9): 1027-1032.

**Mangaleshkar SR, Prasad PS, Chugh S, Thomas AP.** *Staged bilateral total knee replacement--a safer approach in older patients.* Knee. 2001; **8**(3): 207-211.

**Mannion R and Davies HT.** *Reporting health care performance: learning from the past, prospects for the future.* J Eval.Clin.Pract. 2002; **8**(2): 215-228.

**Mant J and Hicks N.** *Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction.* BMJ 1995; **311**(7008): 793-796.

**Marazzi A, Paccaud F, Ruffieux C, Beguin C.** *Fitting the distributions of length of stay by parametric models.* Med Care 1998; **36**(6): 915-927.

**Marcin JP, Slonim AD, Pollack MM, Ruttimann UE.** *Long-stay patients in the pediatric intensive care unit.* Crit Care Med. 2001; **29**(3): 652-657.

**Martin-Baranera M, Planas I, Palau J, Sanz F.** *IMASIS computer-based medical record project: dealing with the human factor.* Medinfo. 1995; **8 Pt 1**: 333.

**Mascioli SR, Jacobs DR, Jr., Kottke TE.** *Diagnostic criteria for hospitalized acute myocardial infarction: the Minnesota experience.* Int.J Epidemiol. 1989; **18**(1): 76-83.

**Maxfield M, Schweitzer J, Gouvier WD.** *Measures of central tendency, variability, and relative standing in nonnormal distributions: alternatives to the mean and standard score.* Arch.Phys.Med Rehabil. 1988; **69**(6): 406-409.

**Mayo NE, Chockalingam A, Reeder BA, Phillips S.** *Surveillance for stroke in Canada.* Health Rep. 1994; **6**(1): 62-72.

**Mayo NE, Danys I, Carlton J, Scott SC.** *Accuracy of hospital discharge coding for stroke.* Can J Cardiol. 1993; **9**(Suppl D): 121D-124D.

**McClellan M and Kessler D.** *A global analysis of technological change in health care: the case of heart attacks. The TECH Investigators.* Health Aff.(Millwood.) 1999; **18**(3): 250-255.

**McCord J.** *Using HIS data for health care research.* World Hosp. 1992; **28**(2): 5-10.

**McLean M, Duclos P, Jacob P, Humphreys P.** *Incidence of Guillain-Barre syndrome in Ontario and Quebec, 1983-1989, using hospital service databases.* Epidemiology 1994; 5(4): 443-448.

**McMahon LF, Jr. and Smits HL.** *Can Medicare prospective payment survive the ICD-9-CM disease classification system?* Ann.Intern.Med. 1986; 104(4): 562-566.

**Melnychuk D, Moride Y, Abenhaim L.** *Monitoring of drug utilization in public health surveillance activities: a conceptual framework.* Can.J.Public Health 1993; 84(1): 45-49.

**Metge C, Black C, Peterson S, Kozyrskyj AL.** *The population's use of pharmaceuticals.* Med Care 1999; 37(6 Suppl): JS42-JS59.

**Miller E, Blatman B, Einarson TR.** *A survey of population-based drug databases in Canada.* CMAJ. 1996; 154(12): 1855-1864.

**Mills TJ, Smith HC, Vlietstra RE.** *PTCA in the elderly: results and expectations.* Geriatrics 1989; 44(7): 71-79.

**Milne R and Clarke A.** *Can readmission rates be used as an outcome indicator?* BMJ 1990; 301(6761): 1139-1140.

**Mitchell JB.** *Using physician claims to supplement hospital data.* In: Grady ML, (editor). *Medical Effectiveness Research Data Methods: Summary Report.* Rockville, MD: Agency for Health Care Policy and Research Pub. No. 92-0056, pp.7 7-81. 1992.

**Mitchell JB, Bubolz T, Paul JE, Pashos CL, Escarce JJ, Muhlbaier LH, Wiesman JM, Young WW, Epstein RS, Javitt JC.** *Using Medicare claims for outcomes research.* Med.Care 1994; 32(7 Suppl): JS38-JS51.

**Mozes B, Schiff E, Modan B.** *Factors affecting inappropriate hospital stay.* Qual.Assur.Health Care 1991; 3(3): 211-217.

**Muhajarine N, Mustard C, Roos LL, Young TK, Gelskey DE.** *Comparison of survey and physician claims data for detecting hypertension.* J.Clin.Epidemiol. 1997; 50(6): 711-718.

**Murphy TF, Sethi S, Niederman MS.** *The role of bacteria in exacerbations of COPD. A constructive view.* Chest 2000; **118**(1): 204-209.

**Mustard CA, Harman CR, Hall PF, Derksen S.** *Impact of a nurses' strike on the cesarean birth rate.* Am.J.Obstet.Gynecol. 1995; **172**(2 Pt 1): 631-637.

**National Institutes of Health.** *National Heart, Lung, and Blood Institute: Morbidity and Mortality: 1996 Chartbook on Cardiovascular, Lung, and Blood Diseases.* 1996; pp 5-9: 49-53.

**Newfoundland Department of Health.** *Report of the Newfoundland reabstracting project.* St-John's, Department of Health 1995.

**Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group.** *Estimation of the incidence of acute myocardial infarction using record linkage: a feasibility study in Nova Scotia and Saskatchewan.* Can J Public Health 1989; **80**: 412-427.

**O'Connor GT, Plume SK, Wennberg JE.** *Regional organization for outcomes research.* Ann.N.Y.Acad.Sci. 1993; **703**: 44-50.

**O'Hara DA and Carson NJ.** *Reporting of adverse events in hospitals in Victoria, 1994-1995.* Med.J Aust. 1997; **166**(9): 460-463.

**Ontario Hospital Association.** *Quality: Report of the Ontario data quality reabstracting study.* Toronto, ON: Ontario Ministry of Health 1991: 1-129.

**Ottenbacher KJ.** *The interpretation of averages in health professions research. An empirical examination.* Eval.Health Prof. 1993; **16**(3): 333-341.

**Pahor M, Carosella L, Pedone C, Manto A, Carbonin P.** *Trends of the characteristics and appropriateness of admissions to acute geriatric and medical wards in Italy from 1988 through 1993. Gruppo Italiano di Farmacovigilanza nell'Anziano--GIFA.* Eur.J Epidemiol. 1996; **12**(6): 563-571.

**Pervez H, Bhargwa A, Parker MJ.** *Accuracy and reliability of the clinical indicators related to hip fractures.* Injury 2003; **34**(7): 522-524.

**Petersen LA, Wright S, Normand SL, Daley J.** *Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database.* J Gen Intern Med 1999; **14**(9): 555-558.

**Petty TL.** *Lung cancer and chronic obstructive pulmonary disease.* Hematol Oncol Clin North Am. 1997; **11**(3): 531-541.

**Pilote L, Lavoie F, Ho V, Eisenberg MJ.** *Changes in the treatment and outcomes of acute myocardial infarction in Quebec, 1988-1995.* CMAJ. 2000; **163**(1): 31-36.

**Pinfold SP, Goel V, Sawka C.** *Quality of hospital discharge and physician data for type of breast cancer surgery.* Med Care 2000; **38**(1): 99-107.

**Pollack MM.** *Cost containment: the pediatric perspective.* New Horiz. 1994; **2**(3): 305-311.

**Pollack MM, Getson PR, Ruttimann UE, Steinhart CM, Kanter RK, Katz RW, Zucker AR, Glass NL, Spohn WA, Fuhrman BP, .** *Efficiency of intensive care. A comparative analysis of eight pediatric intensive care units.* JAMA 1987; **258**(11): 1481-1486.

**Potvin L and Champagne F.** *Utilization of administrative files in health research.* Social Indicators Research 1986; **18**: 409-423.

**Practice Management Information Corporation.** *International Classification of Diseases, 9th Revision. Vol 1-3, Millenium Edition.* Los Angeles 2001.

**Rahme E, Pilote L, LeLorier J.** *Association between naproxen use and protection against acute myocardial infarction.* Arch Intern Med 2002; **162**(10): 1111-1115.

**Ratcliff R.** *Methods for dealing with reaction time outliers.* Psychol Bull. 1993; **114**(3): 510-532.

**Rawson NS and Malcolm E.** *Validity of the recording of cholecystectomy and hysterectomy in the Saskatchewan Health Care datafiles.* Saskatoon: Pharmacoepidemiology Research Consortium 1995a.

**Rawson NS and Malcolm E.** *Validity of the recording of ischaemic heart disease and chronic obstructive pulmonary disease in the Saskatchewan health care datafiles.* Stat.Med 1995b; **14**(24): 2627-2643.

**Rawson NS, Malcolm E, D'Arcy C.** *Reliability of the recording of schizophrenia and depressive disorder in the Saskatchewan health care datafiles.* Soc.Psychiatry Psychiatr.Epidemiol. 1997; **32**(4): 191-199.

**Ray WA.** *Policy and program analysis using administrative databases.* Annals of Internal Medicine 1997; **127**(8 Pt 2): 712-718.

**Ray WA, Griffin MR, Downey W, Melton LJ, III.** *Long-term use of thiazide diuretics and risk of hip fracture.* Lancet 1989; **1**(8640): 687-690.

**Régie de l'Assurance Maladie du Québec.** *Régie de l'assurance maladie du Québec, Statistiques annuelles 1993.* 1994.

**Régie de l'Assurance Maladie du Québec.** *Statistiques Annuelles.* 1996.

**Rich MW, Beckham V, Wittenberg C, Leven CL, Freedland KE, Carney RM.** *A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure.* N.Engl.J Med. 1995; **333**(18): 1190-1195.

**Robinson JR and Tataryn DJ.** *Reliability of the Manitoba Mental Health Management Information System for Research.* Can.J.Psychiatry 1997; **42**(7): 744-749.

**Robinson JR, Young TK, Roos LL, Gelskey DE.** *Estimating the burden of disease. Comparing administrative data and self-reports.* Med Care 1997; **35**(9): 932-947.

**Rodrigues EJ, Simpson E, Richard H, Pilote L.** *Regional variation in the management of acute myocardial infarction in the province of Quebec.* Can J Cardiol. 2002; **18**(10): 1067-1076.

**Romano PS.** *Can administrative data be used to compare the quality of health care?* Med.Care Rev. 1993a; **50**(4): 451-477.

**Romano PS and Mark DH.** *Bias in the coding of hospital discharge data and its implications for quality assessment.* Med Care 1994; **32**(1): 81-90.

**Romano PS, Roos LL, Jollis JG.** *Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives.* J.Clin.Epidemiol. 1993b; 46(10): 1075-1079.

**Romano PS, Roos LL, Luft HS, Jollis JG, Doliszny K.** *A comparison of administrative versus clinical data: coronary artery bypass surgery as an example. Ischemic Heart Disease Patient Outcomes Research Team.* J Clin.Epidemiol. 1994; 47(3): 249-260.

**Roos LL and Brazauskas R.** *Outcomes and quality assurance: facilitating the use of administrative data.* Qual.Assur.Health Care 1990; 2(1): 77-88.

**Roos LL, Fisher ES, Sharp SM, Newhouse JP, Anderson G, Bubolz TA.** *Postsurgical mortality in Manitoba and New England.* JAMA 1990; 263(18): 2453-2458.

**Roos LL, Mustard CA, Nicol JP, McLerran DF, Malenka DJ, Young TK, Cohen MM.** *Registries and administrative data: organization and accuracy.* Med.Care 1993; 31(3): 201-212.

**Roos LL, Roos NP, Cageorge SM, Nicol JP.** *How good are the data? Reliability of one health care data bank.* Med.Care 1982; 20(3): 266-276.

**Roos LL and Sharp SM.** *Becoming more efficient at outcomes research.* Int.J.Technol.Assess.Health Care 1988; 4(4): 555-571.

**Roos LL, Sharp SM, Cohen MM.** *Comparing clinical information with claims data: some similarities and differences.* J.Clin.Epidemiol. 1991; 44(9): 881-888.

**Roos LL, Sharp SM, Wajda A.** *Assessing data quality: a computerized approach.* Soc.Sci.Med. 1989; 28(2): 175-182.

**Roos LL, Stranc L, James RC, Li J.** *Complications, comorbidities, and mortality: improving classification and prediction.* Health Serv.Res. 1997; 32(2): 229-238.

**Roos LL, Walld R, Wajda A, Bond R, Hartford K.** *Record linkage strategies, outpatient procedures, and administrative data.* Med Care 1996a; 34(6): 570-582.

**Roos LL, Walld RK, Romano PS, Roberecki S.** *Short-term mortality after repair of hip fracture. Do Manitoba elderly do worse?* Med.Care 1996b; 34(4): 310-326.

**Roos LLJr, Nicol JP, Cageorge SM.** *Using administrative data for longitudinal research: comparisons with primary data collection.* J.Chronic.Dis. 1987; 40(1): 41-49.

**Roos LLJr, Roos NP, Cageorge SM, Nicol JP.** *How good are the data? Reliability of one health care data bank.* Med.Care 1982; 20(3): 266-276.

**Roos NP, Black C, Frohlich N, DeCoster C, Cohen M, Tataryn DJ, Mustard CA, Roos LL, Toll F, Carriere KC, Burchill CA, MacWilliam L, Bogdanovic B.** *Population health and health care use: an information system for policy makers.* Milbank Q. 1996; 74(1): 3-31.

**Roos NP, Black CD, Frohlich N, DeCoster C, Cohen MM, Tataryn DJ, Mustard CA, Toll F, Carriere KC, Burchill CA, .** *A population-based health information system.* Med.Care 1995; 33(12 Suppl): DS13-DS20.

**Roos NP, Flowerdew G, Wajda A, Tate RB.** *Variations in physicians' hospitalization practices: a population-based study in Manitoba, Canada.* Am.J Public Health 1986; 76(1): 45-51.

**Roos NP, Roos LL, Mossey J, Havens B.** *Using administrative data to predict important health outcomes. Entry to hospital, nursing home, and death.* Med.Care 1988; 26(3): 221-239.

**Ruttimann UE, Patel KM, Pollack MM.** *Length of stay and efficiency in pediatric intensive care units.* J.Pediatr. 1998; 133(1): 79-85.

**Ruttimann UE and Pollack MM.** *Variability in duration of stay in pediatric intensive care units: a multiinstitutional study.* J.Pediatr. 1996; 128(1): 35-44.

**Sackett DL, Haynes RB, Guyatt GH, Tugwell P.** *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Boston, MA: Little, Brown and Company 1991: p.1-441.

**Sackett DL.** *Bias in analytic research.* J Chronic.Dis 1979; 32(1-2): 51-63.



**Samuels ML and Witmer JA.** *Statistics for the life sciences. 2nd ed.* 1999.

**Sanazaro PJ and Mills DH.** *A critique of the use of generic screening in quality assessment.* JAMA 1991; **265**(15): 1977-1981.

**Sandford AJ and Pare PD.** *Genetic risk factors for chronic obstructive pulmonary disease.* Clin.Chest Med. 2000; **21**(4): 633-643.

**Schellevis FG, van de LE, van d, V, van Eijk JT, van Weel C.** *Validity of diagnoses of chronic diseases in general practice. The application of diagnostic criteria.* J Clin.Epidemiol. 1993; **46**(5): 461-468.

**Scherer K, Kraut A, Yassi A, Wajda A, Bebhuk J.** *Using administrative health data to monitor potential adverse health effects in environmental studies.* Environ.Res. 1994; **66**(2): 143-151.

**Schiff GD and Yaacoub AS.** *The diagnostic coding of myocardial infarction.* Annals of Internal Medicine 1989; **110**(3): 243.

**Schulman LL.** *Lung transplantation for chronic obstructive pulmonary disease.* Clin.Chest Med. 2000; **21**(4): 849-865.

**Seagroatt V, Tan HS, Goldacre M, Bulstrode C, Nugent I, Gill L.** *Elective total hip replacement: incidence, emergency readmission rate, and postoperative mortality.* BMJ 1991; **303**(6815): 1431-1435.

**Sethi S.** *Bacterial infection and the pathogenesis of COPD.* Chest 2000; **117**(5 Suppl 1): 286S-291S.

**Shaheen SO, Sterne JA, Tucker JS, Florey CD.** *Birth weight, childhood lower respiratory tract infection, and adult lung function.* Thorax 1998; **53**(7): 549-553.

**Sheehy O and LeLorier J.** *Patterns of amlodipine and felodipine use in an elderly Quebec population.* Can J Cardiol. 2000; **16**(9): 1109-1117.

**Sheiner LB and Beal SL.** *Some suggestions for measuring predictive performance.* J Pharmacokinet.Biopharm. 1981; **9**(4): 503-512.

**Sherk PA and Grossman RF.** *The chronic obstructive pulmonary disease exacerbation.* Clin.Chest Med. 2000; **21**(4): 705-721.

**Sisk JE, Dougherty DM, Ehrenhaft PM, Ruby G, Mitchner BA.** *Assessing information for consumers on the quality of medical care.* Inquiry 1990; **27**(3): 263-272.

**Smith DM, Norton JA, McDonald CJ.** *Nonelective readmissions of medical patients.* J.Chronic.Dis. 1985; **38**(3): 213-224.

**Snow V, Lascher S, Mottur-Pilson C.** *The evidence base for management of acute exacerbations of COPD: clinical practice guideline, part 1.* Chest 2001; **119**(4): 1185-1189.

**Solin LJ, Legorreta A, Schultz DJ, Levin HA, Zatz S, Goodman RL.** *Analysis of a claims database for the identification of patients with carcinoma of the breast.* J Med.Syst. 1994; **18**(1): 23-32.

**Soofi ES.** *Capturing the intangible concept of information.* J Am Stat Assoc. 1994; **89**: 1243-1254.

**Sorensen HT, Sabroe S, Olsen J.** *A framework for evaluation of secondary data sources for epidemiological research.* Int.J.Epidemiol. 1996; **25**(2): 435-442.

**Soto FJ and Varkey B.** *Evidence-based approach to acute exacerbations of COPD.* Curr.Opin.Pulm.Med. 2003; **9**(2): 117-124.

**South M.** *Reduction in length of hospital stay for acute childhood asthma associated with the introduction of casemix funding.* Med.J Aust. 1997; **167**(1): 11-13.

**Sperry S and Birdsall C.** *Outcomes of a pneumonia critical path.* Nurs.Econ. 1994; **12**(6): 332-9, 345.

**Statistics Canada.** *Canadian Classification of Diagnostic, Therapeutic and Surgical Procedures, 2nd Edition .* 1986: Ministry of Supply and Services Canada.

**Steinwachs DM, Stuart ME, Scholle S, Starfield B, Fox MH, Weiner JP.** *A comparison of ambulatory Medicaid claims to medical records: a reliability assessment.* Am.J Med Qual. 1998; **13**(2): 63-69.

**Stewart S, Pearson S, Horowitz JD.** *Effects of a home-based intervention among patients with congestive heart failure discharged from acute hospital care.* Arch.Intern.Med. 1998; **158**(10): 1067-1072.

**Stokes J and Lindsay J.** *Major causes of death and hospitalization in Canadian seniors.* Chronic.Dis.Can. 1996; **17**(2): 63-73.

**Sweet L, Fan JSK, O'Neill J.** *A validation of asthma, Prince Edward Island, 1984-1988: II. Hospitalization.* Chronic.Dis.Can. 1992; **13**(2): 19-23.

**Szklo.M and Nieto FJ.** *Epidemiology: Beyond the Basics.* 2000: 362-404.

**Tamblyn R, Lavoie G, Petrella L, Monette J.** *The use of prescription claims databases in pharmacoepidemiological research: the accuracy and comprehensiveness of the prescription claims database in Quebec.* J.Clin.Epidemiol. 1995; **48**(8): 999-1009.

**Tamblyn R, Reid T, Mayo N, McLeod P, Churchill-Smith M.** *Using medical services claims to assess injuries in the elderly: sensitivity of diagnostic and procedure codes for injury ascertainment.* J.Clin.Epidemiol. 2000; **53**(2): 183-194.

**Tennis P, Bombardier C, Malcolm E, Downey W.** *Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan Hospital Separations Database.* J.Clin.Epidemiol. 1993; **46**(7): 675-683.

**Testa MA and Simonson DC.** *Assesment of quality-of-life outcomes.* N.Engl.J.Med. 1996; **334**(13): 835-840.

**The Doctor's Hospital.** *Data quality and utilization study.* Toronto, The Doctor's Hospital 1992.

**The Quality Indicator Study Group.** *An approach to the evaluation of quality indicators of the outcome of care in hospitalized patients, with a focus on nosocomial infection indicators.* Infect.Control Hosp.Epidemiol. 1995; **16**(5): 308-316.

**Thiessen BQ, Wallace SM, Blackburn JL, Wilson TW, Bergman U.** *Increased prescribing of antidepressants subsequent to beta-blocker therapy.* Arch.Intern Med 1990; **150**(11): 2286-2290.

**Tsang P and Severs MP.** *A study of appropriateness of acute geriatric admissions and an assessment of the Appropriateness Evaluation Protocol.* J R.Coll.Physicians Lond 1995; **29**(4): 311-314.

**Tu JV, Mazer CD, Levinton C, Armstrong PW, Naylor CD.** *A predictive index for length of stay in the intensive care unit following cardiac surgery.* CMAJ. 1994; **151**(2): 177-185.

**Tunstall-Pedoe H.** *The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. WHO MONICA Project Principal Investigators.* J Clin.Epidemiol. 1988; **41**(2): 105-114.

**van Walraven C, Wang B, Ugnat AM, Naylor CD.** *False-positive coding for acute myocardial infarction on hospital discharge records: chart audit results from a tertiary centre.* Can J Cardiol. 1990; **6**(9): 383-386.

**Waite K, Oddone E, Weinberger M, Samsa G, Foy M, Henderson W.** *Lack of association between patients' measured burden of disease and risk for hospital readmission.* J.Clin.Epidemiol. 1994; **47**(11): 1229-1236.

**Walter SD and Cook RJ.** *A comparison of several point estimators of the odds ratio in a single 2 x 2 contingency table.* Biometrics 1991; **47**(3): 795-811.

**Wang FL, Gabos S, Sibbald B, Lowry RB.** *Completeness and accuracy of the birth registry data on congenital anomalies in Alberta, Canada.* Chronic.Dis.Can 2001; **22**(2): 57-66.

**Weinberger M, Oddone EZ, Henderson WG.** *Does increased access to primary care reduce hospital readmissions? Veterans Affairs Cooperative Study Group on Primary Care and Hospital Readmission.* N.Engl.J Med. 1996; **334**(22): 1441-1447.

**Weingarten S, Riedinger MS, Sandhu M, Bowers C, Ellrodt AG, Nunn C, Hobson P, Greengold N.** *Can practice guidelines safely reduce hospital length of stay? Results from a multicenter interventional study.* Am.J.Med. 1998; **105**(1): 33-40.

**Wennberg J and Gittelsohn.** *Small area variations in health care delivery.* Science 1973; **182**(117): 1102-1108.

**West SL, Richter A, Melfi CA, McNutt M, Nennstiel ME, Mauskopf JA.** *Assessing the Saskatchewan database for outcomes research studies of depression and its treatment.* J.Clin.Epidemiol. 2000; 53(8): 823-831.

**White E.** *The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria.* Am.J.Epidemiol. 1986; 124(5): 816-825.

**Williams JI and Young W.** *A summary of studies on the quality of health care administrative databases in Canada. In Goel V et al. (eds). Patterns of Health Care in Ontario. The ICES Practical Atlas. 2nd edition. Ottawa, ON: Canadian Medical Association: 339-3345. 1996a.*

**Williams JI and Young W.** *Inventory of Studies on the Accuracy of Canadian Health Administrative Databases.* ICES, North York. 1996b.

**Wolff N and Helminiak TW.** *Nonsampling measurement error in administrative data: implications for economic evaluations.* Health Econ. 1996; 5(6): 501-512.

**Wray NP, Ashton CM, Kuykendall DH, Hollingsworth JC.** *Using administrative databases to evaluate the quality of medical care: a conceptual framework.* Soc.Sci.Med. 1995; 40(12): 1707-1715.

**Wu G.** *Calculating predictive performance: a user's note.* Pharmacol.Res. 1995; 31(6): 393-399.

**Young TK, Roos NP, Hammerstrand KM.** *Estimated burden of diabetes mellitus in Manitoba according to health insurance claims: a pilot study.* CMAJ. 1991; 144(3): 318-324.

# Appendix 1

**Table 1A - Statistics for the assessment of diagnostic tests [Mackinnon, 2000].**

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)	Test				
		Positive	Negative		
		370	35	405	
	Positive	370	35	405	
	Negative	25	70	95	
		395	105	500	
Confidence Intervals					
95%					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	91.36%	1.40%	88.19%	93.91%
Sensitivity of a random test	SE_RAN	79.00%			
Quality index of sensitivity	$\kappa(1,0)$	58.85%			
Specificity	SP	73.68%	4.52%	63.65%	82.19%
Specificity of a random test	SP_RAN	21.00%			
Quality index of specificity	$\kappa(0,0)$	66.96%			
Efficiency (Correct classification rate)	EFF	88.00%	1.45%	84.82%	90.72%
Efficiency of a random test	EFF_RAN	67.98%			
Quality index	$\kappa(1,0)$	62.52%	4.40%	53.90%	71.15%
Youden's Index	J	85.04%	4.73%	55.77%	74.31%
Predictive value of positive test	PVP	93.67%	1.23%	90.80%	95.86%
Pred. value of a positive random test	PVP_RAN	81.00%			
Predictive value of negative test	PVN	66.67%	4.80%	56.80%	75.57%
Pred. value of a negative random test	PVN_RAN	19.00%			
False positive rate	FP	26.32%	4.52%	17.81%	36.35%
False negative rate	FN	8.64%	1.40%	6.09%	11.81%
Misclassification rate	1-EFF	12.00%	1.45%	9.28%	17.02%
Prevalence	P	81.00%	1.75%	77.28%	84.35%
Test level	Q	79.00%	1.82%	75.16%	82.49%
Likelihood ratio of positive test	LR+	3.4716	1.1881	2.4764	4.8668
Likelihood ratio of negative test	LR-	8.5263	1.1886	6.0767	11.9634
Odds ratio	OR	29.6000	1.3398	16.8844	52.5138
Odds ratio (Haldane's estimator)	OR*	28.8542	1.3368	16.3348	50.9687
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.6252	0.0440	0.5390	0.7115
This kappa indicates substantial agreement. Test of Ho: Kappa=0: z=14.01, p=0.0000 L.L.					
Observed Agreement	PO	0.8800	0.0145	0.8482	0.9072
Chance Agreement	PE	0.6798			
Positive Agreement	PA	0.9250	0.0087	0.9061	0.9439
Negative Agreement	NA	0.7000	0.0369	0.6276	0.7724
Byrt's Bias Index	BI	0.0200			
Byrt's Prevalence Asymmetry Index	PI	-0.6000			
Bias Adjusted Kappa	BAK	0.6250			
Prevalence & Bias Adjusted Kappa	PABAK	0.7600			
Alternative Indices of Association					
Dice's Index	p(s)	0.9250	0.0097	0.9061	0.9439
Yule's Q (Gamma)	$\gamma$	0.9346	0.0185	0.8984	0.9709
Pearson Correlation (Phi)	$\phi$	0.6265	0.0439	0.5405	0.7124
Scott's agreement index	$\pi$	0.6250			
Tetrachoric Correlation	r(t)	0.8834	0.0241	0.8362	0.9306
Goodman & Kruskal's tau (Crit. dep.)	$\tau(asy)$	0.3924	0.0550	0.2847	0.5002
Lambda(Symmetric)	$\lambda(sym)$	0.4000	0.0739	0.2552	0.5448
Lambda(Criterion dep.)	$\lambda(asy)$	0.3684	0.0857	0.2004	0.5364
Uncertainty Coefficient (Symmetric)	U(sym)	0.3323	0.0468	0.2405	0.4241
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.3418	0.0481	0.2475	0.4361
Chi Square					
Test	Chi Square	P			
Pearson Chi Square	196.22	0.0000			
with Yate's correction	182.32	0.0000			
Likelihood Ratio Chi Square	166.17	0.0000			
Minimum Expected Frequency	19.95				
Cells with Expected Frequency < 5	0 of 4 (0%)				
Cells with Expected Frequency < 1	0 of 4 (0%)				
McNemar's Test					
Test	Chi Square	P			
McNemar's Test	1.67	0.1967			
with Yate's correction	1.35	0.2453			
Kullback-Leibler Distance					
D(f  g)	1.32227				
D(g  f)	-1.90670				
Pin	3.75192				
Pout	0.14857				

	Test		
	Positive	Negative	
	Row 91.36% Col 93.67%	Row 8.64% Col 33.33%	81.00%
Positive	Row 26.32% Col 6.33%	Row 73.68% Col 66.67%	19.00%
Negative	79.00%	21.00%	N = 500

No. Decimal Places

Summary  
Sensitivity = 0.9136 (95% CI: 0.8819 - 0.9391)

Specificity = 0.7368 (95% CI: 0.6365 - 0.8219)

Efficiency (Correct classification rate) = 0.8800 (95% CI: 0.8482 - 0.9072)

Quality index = 0.6252 (95% CI: 0.5390 - 0.7115)

Youden's index = 0.6504 (95% CI: 0.5577 - 0.7431)

Predictive value of positive test = 0.9367 (95% CI: 0.9080 - 0.9586)

Predictive value of negative test = 0.6667 (95% CI: 0.5680 - 0.7557)

False positive rate = 0.2632 (95% CI: 0.1781 - 0.3635)

False negative rate = 0.0864 (95% CI: 0.0609 - 0.1181)

Misclassification rate = 0.1200 (95% CI: 0.0928 - 0.1702)

Prevalence = 0.8100 (95% CI: 0.7728 - 0.8435)

Test level = 0.7900 (95% CI: 0.7516 - 0.8249)

Likelihood ratio of positive test = 3.4716 (95% CI: 2.4764 - 4.8668)

Likelihood ratio of negative test = 8.5263 (95% CI: 6.0767 - 11.9634)

Odds ratio = 29.6000 (95% CI: 16.8844 - 52.5138)

Odds ratio (Haldane's estimator) = 28.8542 (95% CI: 16.3348 - 50.9687)

Cohen's Kappa = 0.6252 (95% CI: 0.5390 - 0.7115)

This kappa indicates substantial agreement.

Test of Ho: Kappa=0: z=14.01, p=0.0000 L.L.

Observed Agreement = 0.8800 (95% CI: 0.8482 - 0.9072)

Chance Agreement = 0.6798 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.9250 (95% CI: 0.9061 - 0.9439)

Negative Agreement = 0.7000 (95% CI: 0.6276 - 0.7724)

Dice's Index = 0.9250 (95% CI: 0.9061 - 0.9439)

Yule's Q (Gamma) = 0.9346 (95% CI: 0.8984 - 0.9709)

Pearson Correlation (Phi) = 0.6265 (95% CI: 0.5405 - 0.7124)

Scott's agreement index = 0.6250

Tetrachoric Correlation = 0.8834 (95% CI: 0.8362 - 0.9306)

Goodman & Kruskal's tau (Crit. dep.) = 0.3924 (95% CI: 0.2847 - 0.5002)

Lambda(Symmetric) = 0.4000 (95% CI: 0.2552 - 0.5448)

Lambda(Criterion dep.) = 0.3684 (95% CI: 0.2004 - 0.5364)

Uncertainty Coefficient (Symmetric) = 0.3323 (95% CI: 0.2405 - 0.4241)

Uncertainty Coeff. (Criterion dep.) = 0.3418 (95% CI: 0.2475 - 0.4361)

Pearson Chi Square = 196.2245, df=1, p=0.000000

Pearson Chi Square with Yate's correction = 192.3235, df=1, p=0.000000

Likelihood Ratio Chi Square = 166.1713, df=1, p=0.000000

McNemar's Test = 1.6667, df=1, p=0.196706

McNemar's Test with Yate's correction = 1.3500, df=1, p=0.245278

# Appendix 2

**Table 2A – Diagnostic and agreement statistics for the COPD cohort.**

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)	Positive Negative	Test		1206 2562 3768	
		Positive	Negative		
		1115	91		
		118	2444		
		1233	2535		
Confidence Intervals					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	92.45%	0.76%	90.82%	93.88%
Sensitivity of a random test	SE_RAN	32.72%			
Quality Index of sensitivity	$\kappa(1,0)$	88.78%			
Specificity	SP	95.30%	0.41%	94.51%	96.17%
Specificity of a random test	SP_RAN	67.28%			
Quality Index of specificity	$\kappa(0,0)$	85.82%			
Efficiency (Correct classification rate)	EFF	94.45%	0.37%	93.67%	95.16%
Efficiency of a random test	EFF_RAN	58.22%			
Quality Index	$\kappa(5,0)$	87.33%	0.85%	85.66%	89.00%
Youden's Index	J	87.85%	0.87%	86.15%	89.55%
Predictive value of positive test	PVP	90.43%	0.84%	88.65%	92.01%
Pred. value of a positive random test	PVP_RAN	32.01%			
Predictive value of negative test	PVN	96.41%	0.37%	95.61%	97.10%
Pred. value of a negative random test	PVN_RAN	67.99%			
False positive rate	FP	4.61%	0.41%	3.83%	5.49%
False negative rate	FN	7.55%	0.76%	6.12%	9.18%
Misclassification rate	1-EFF	5.55%	0.37%	4.84%	5.80%
Prevalence	P	32.01%	0.76%	30.52%	33.52%
Test level	Q	32.72%	0.76%	31.23%	34.25%
Likelihood ratio of positive test	LR+	20.0736	1.0945	16.8179	23.9595
Likelihood ratio of negative test	LR-	12.6424	1.1062	10.3741	15.4065
Odds ratio	OR	253.7772	1.1550	191.3288	336.6084
Odds ratio (Haldane's estimator)	OR*	251.4897	1.1546	189.7285	333.3556
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.8733	0.0085	0.8566	0.8900
This kappa indicates almost perfect agreement. Test of Ho: Kappa=0: z=53.61, p =0.0000 t.t.t.					
Observed Agreement	PO	0.9445	0.0037	0.9367	0.9516
Chance Agreement	PE	0.5622			
Positive Agreement	PA	0.9143	0.0059	0.9027	0.9259
Negative Agreement	NA	0.9590	0.0028	0.9534	0.9645
Byrt's Bias Index	BI	-0.0072			
Byrt's Prevalence Asymmetry Index	PI	0.3527			
Bias Adjusted Kappa	BAK	0.8733			
Prevalence & Bias Adjusted Kappa	PABAK	0.8891			
Alternative Indices of Association					
Dice's Index	p(s)	0.9143	0.0059	0.9027	0.9259
Yule's Q (Gamma)	$\gamma$	0.9922	0.0011	0.9899	0.9944
Pearson Correlation (Phi)	$\phi$	0.8734	0.0085	0.8568	0.8901
Scott's agreement index	$\kappa$	0.8733			
Tetrachoric Correlation	r(t)	0.9828	0.0018	0.9792	0.9864
Goodman & Kruskal's tau (Crit. dep.)	$\tau(asy)$	0.7629	0.0148	0.7338	0.7920
Lambda(Symmetric)	$\lambda(sym)$	0.8286	0.0118	0.8055	0.8518
Lambda(Criterion dep.)	$\lambda(asy)$	0.8267	0.0121	0.8029	0.8505
Uncertainty Coefficient (Symmetric)	U(sym)	0.6665	0.0163	0.6346	0.6984
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.6693	0.0163	0.6374	0.7012
Chi Square					
Test	Chi Square	P			
Pearson Chi Square	2874.53	0.0000			
with Yate's correction	2870.54	0.0000			
Likelihood Ratio Chi Square	3162.12	0.0000			
Minimum Expected Frequency	394.64				
Cells with Expected Frequency < 5	0 of 4 (0%)				
Cells with Expected Frequency < 1	0 of 4 (0%)				
McNemar's Test					
Test	Chi Square	P			
McNemar's Test	3.49	0.0618			
with Yate's correction	3.23	0.0721			
Kullback-Leibler Distance					
D(I  G)	2.96452				
D(G  I)	-2.55835				
Pin	19.38534				
Pout	0.07743				

	Test		32.01%
	Positive	Negative	
	Row 92.45% Col 92.43%	Row 7.55% Col 3.9%	
Positive			
Negative	Row 4.61% Col 9.57%	Row 95.39% Col 96.41%	67.99%
		32.72%	67.28%
		N = 3768	

No. Decimal Places

Summary  
Sensitivity = 0.9245 (95% CI: 0.9082 - 0.9388)

Specificity = 0.9539 (95% CI: 0.9451 - 0.9617)

Efficiency (Correct classification rate) = 0.9445 (95% CI: 0.9367 - 0.9516)

Quality index = 0.8733 (95% CI: 0.8566 - 0.8900)

Youden's index = 0.8785 (95% CI: 0.8615 - 0.8955)

Predictive value of positive test = 0.9043 (95% CI: 0.8865 - 0.9201)

Predictive value of negative test = 0.9641 (95% CI: 0.9561 - 0.9710)

False positive rate = 0.0461 (95% CI: 0.0383 - 0.0549)

False negative rate = 0.0755 (95% CI: 0.0612 - 0.0918)

Misclassification rate = 0.0555 (95% CI: 0.0484 - 0.0560)

Prevalence = 0.3201 (95% CI: 0.3052 - 0.3352)

Test level = 0.3272 (95% CI: 0.3123 - 0.3425)

Likelihood ratio of positive test = 20.0736 (95% CI: 16.8179 - 23.9595)

Likelihood ratio of negative test = 12.6424 (95% CI: 10.3741 - 15.4065)

Odds ratio = 253.7772 (95% CI: 191.3288 - 336.6084)

Odds ratio (Haldane's estimator) = 251.4897 (95% CI: 189.7285 - 333.3556)

Cohen's Kappa = 0.8733 (95% CI: 0.8566 - 0.8900)

This kappa indicates almost perfect agreement.

Test of Ho: Kappa=0: z=53.61, p=0.0000 t.t.t.

Observed Agreement = 0.9445 (95% CI: 0.9367 - 0.9516)

Chance Agreement = 0.5622 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.9143 (95% CI: 0.9027 - 0.9259)

Negative Agreement = 0.9590 (95% CI: 0.9534 - 0.9645)

Dice's Index = 0.9143 (95% CI: 0.9027 - 0.9259)

Yule's Q (Gamma) = 0.9922 (95% CI: 0.9899 - 0.9944)

Pearson Correlation (Phi) = 0.8734 (95% CI: 0.8568 - 0.8901)

Scott's agreement index = 0.8733

Tetrachoric Correlation = 0.9828 (95% CI: 0.9792 - 0.9864)

Goodman & Kruskal's tau (Crit. dep.) = 0.7629 (95% CI: 0.7338 - 0.7920)

Lambda(Symmetric) = 0.8286 (95% CI: 0.8055 - 0.8518)

Lambda(Criterion dep.) = 0.8267 (95% CI: 0.8029 - 0.8505)

Uncertainty Coefficient (Symmetric) = 0.6665 (95% CI: 0.6346 - 0.6984)

Uncertainty Coeff. (Criterion dep.) = 0.6693 (95% CI: 0.6374 - 0.7012)

Pearson Chi Square = 2874.5254, df=1, p=0.000000

Pearson Chi Square with Yate's correction = 2870.5364, df=1, p=0.000000

Likelihood Ratio Chi Square = 3162.1186, df=1, p=0.000000

McNemar's Test = 3.4880, df=1, p=0.061814

McNemar's Test with Yate's correction = 3.2344, df=1, p=0.072105

# Appendix 3

**Table 3A - Diagnostic and agreement statistics for AMI.**

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)	Positive	Test		19076	688601
		Positive	Negative		
		15468	3608		
		8153	680448		
	Negative	23621	684056	707677	
Confidence Intervals					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	81.09%	0.28%	80.52%	81.64%
Sensitivity of a random test	SE_RAN	3.34%			
Quality index of sensitivity	$\kappa(1,0)$	80.43%			
Specificity	SP	98.82%	0.01%	98.79%	98.84%
Specificity of a random test	SP_RAN	98.66%			
Quality index of specificity	$\kappa(0,0)$	84.53%			
Efficiency (Correct classification rate)	EFF	98.34%	0.02%	98.31%	98.37%
Efficiency of a random test	EFF_RAN	94.15%			
Quality index	$\kappa(5,0)$	71.61%	0.25%	71.12%	72.10%
Youden's Index	J	79.90%	0.28%	79.35%	80.48%
Predictive value of positive test	PVP	65.48%	0.31%	64.87%	66.09%
Prad. value of a positive random test	PVP_RAN	2.70%			
Predictive value of negative test	PVN	99.47%	0.01%	99.46%	99.49%
Prad. value of a negative random test	PVN_RAN	97.30%			
False positive rate	FP	1.18%	0.01%	1.16%	1.21%
False negative rate	FN	18.91%	0.28%	18.36%	19.48%
Misclassification rate	1-EFF	1.66%	0.02%	1.63%	1.05%
Prevalence	P	2.70%	0.02%	2.66%	2.73%
Test level	Q	3.34%	0.02%	3.30%	3.38%
Likelihood ratio of positive test	LR+	68.4853	1.0116	66.9522	70.0534
Likelihood ratio of negative test	LR-	5.2245	1.0151	5.0733	5.3803
Odds ratio	OR	357.8040	1.0218	342.9821	373.2663
Odds ratio (Haldane's estimator)	OR*	357.7443	1.0218	342.8257	373.2032
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.7161	0.0025	0.7112	0.7210
This kappa indicates substantial agreement.					
Test of Ho: Kappa=0: z=606.05, p=0.0000 L.L.L.					
Observed Agreement	PO	0.9834	0.0002	0.9831	0.9837
Chance Agreement	PE	0.9415			
Positive Agreement	PA	0.7245	0.0024	0.7198	0.7293
Negative Agreement	NA	0.9914	0.0001	0.9913	0.9916
Byrt's Bias Index	BI	-0.0064			
Byrt's Prevalence Asymmetry Index	PI	0.9397			
Bias Adjusted Kappa	BAK	0.7160			
Prevalence & Bias Adjusted Kappa	PABAK	0.9668			
Alternative Indices of Association					
Dice's Index	p(s)	0.7245	0.0024	0.7198	0.7293
Yule's Q (Gamma)	$\gamma$	0.9944	0.0001	0.9942	0.9947
Pearson Correlation (Phi)	$\phi$	0.7204	0.0024	0.7156	0.7252
Scott's agreement index	$\kappa$	0.7160			
Tetrachoric Correlation	r(t)	0.9876	0.0002	0.9872	0.9880
Goodman & Kruskal's tau (Crit. dep.)	$\tau(\text{asy})$	0.5190	0.0035	0.5121	0.5259
Lambda(Symmetric)	$\lambda(\text{sym})$	0.4491	0.0049	0.4395	0.4587
Lambda(Criterion dep.)	$\lambda(\text{asy})$	0.3835	0.0063	0.3711	0.3959
Uncertainty Coefficient (Symmetric)	U(sym)	0.5229	0.0030	0.5169	0.5288
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.5699	0.0032	0.5636	0.5761
Test					
Test	Chi Square	P			
Pearson Chi Square	367297.05	0.0000			
with Yates's correction	367272.28	0.0000			
Likelihood Ratio Chi Square	100010.83	0.0000			
Minimum Expected Frequency 636.72					
Cells with Expected Frequency < 5 0 of 4 (0%)					
Cells with Expected Frequency < 1 0 of 4 (0%)					
Test					
Test	Chi Square	P			
McNemar's Test	1756.40	0.0000			
with Yates's correction	1755.63	0.0000			
Kullback-Leibler Distance					
D(f  g)	3.73992				
D(g  f)	-1.68383				
Pin	42.09455				
Pout	0.18566				

Test			
Positive	Positive	Negative	2.70%
	Row 81.09% Col 65.48%	Row 18.91% Col 0.53%	
	Row 1.18% Col 34.52%	Row 98.82% Col 99.47%	
Negative	3.34%	96.66%	97.30%
N = 707677			

No. Decimal Places 4

Summary	
Sensitivity = 0.8109 (95% CI: 0.8052 - 0.8164)	
Specificity = 0.9882 (95% CI: 0.9879 - 0.9884)	
Efficiency (Correct classification rate) = 0.9834 (95% CI: 0.9831 - 0.9837)	
Quality index = 0.7161 (95% CI: 0.7112 - 0.7210)	
Youden's index = 0.7990 (95% CI: 0.7935 - 0.8046)	
Predictive value of positive test = 0.6548 (95% CI: 0.6487 - 0.6609)	
Predictive value of negative test = 0.9947 (95% CI: 0.9946 - 0.9949)	
False positive rate = 0.0118 (95% CI: 0.0116 - 0.0121)	
False negative rate = 0.1891 (95% CI: 0.1836 - 0.1948)	
Misclassification rate = 0.0166 (95% CI: 0.0163 - 0.0105)	
Prevalence = 0.0270 (95% CI: 0.0266 - 0.0273)	
Test level = 0.0334 (95% CI: 0.0330 - 0.0338)	
Likelihood ratio of positive test = 68.4853 (95% CI: 66.9522 - 70.0534)	
Likelihood ratio of negative test = 5.2245 (95% CI: 5.0733 - 5.3803)	
Odds ratio = 357.8040 (95% CI: 342.9821 - 373.2663)	
Odds ratio (Haldane's estimator) = 357.7443 (95% CI: 342.9257 - 373.2032)	
Cohen's Kappa = 0.7161 (95% CI: 0.7112 - 0.7210)	
This kappa indicates substantial agreement.	
Test of Ho: Kappa=0: z=606.05, p=0.0000 L.L.L.	
Observed Agreement = 0.9834 (95% CI: 0.9831 - 0.9837)	
Chance Agreement = 0.9415 (95% CI: 0.0000 - 0.0000)	
Positive Agreement = 0.7245 (95% CI: 0.7198 - 0.7293)	
Negative Agreement = 0.9914 (95% CI: 0.9913 - 0.9916)	
Byrt's Bias Index = -0.0064	
Byrt's Prevalence Asymmetry Index = 0.9397	
Bias Adjusted Kappa = 0.7160	
Prevalence & Bias Adjusted Kappa = 0.9668	
Dice's Index = 0.7245 (95% CI: 0.7198 - 0.7293)	
Yule's Q (Gamma) = 0.9944 (95% CI: 0.9942 - 0.9947)	
Pearson Correlation (Phi) = 0.7204 (95% CI: 0.7156 - 0.7252)	
Scott's agreement index = 0.7160	
Tetrachoric Correlation = 0.9876 (95% CI: 0.9872 - 0.9880)	
Goodman & Kruskal's tau (Crit. dep.) = 0.5190 (95% CI: 0.5121 - 0.5259)	
Lambda(Symmetric) = 0.4491 (95% CI: 0.4395 - 0.4587)	
Lambda(Criterion dep.) = 0.3835 (95% CI: 0.3711 - 0.3959)	
Uncertainty Coefficient (Symmetric) = 0.5229 (95% CI: 0.5169 - 0.5288)	
Uncertainty Coeff. (Criterion dep.) = 0.5699 (95% CI: 0.5636 - 0.5761)	
Pearson Chi Square = 367297.0485, df=1, p=0.000000	
Pearson Chi Square with Yates's correction = 367272.2839, df=1, p=0.000000	
Likelihood Ratio Chi Square = 100010.8292, df=1, p=0.000000	
McNemar's Test = 1756.4004, df=1, p=0.000000	
McNemar's Test with Yates's correction = 1755.6276, df=1, p=0.000000	

	Test		2.70%
	Positive	Negative	
	Row 81.09% Col 65.48%	Row 18.91% Col 0.53%	
Positive	15468	3608	97.30%
Negative	8153	680448	
	3.34%	96.66%	N = 707677

No. Decimal Places

Summary

Sensitivity = 0.8109 (95% CI: 0.8052 - 0.8164)

Specificity = 0.9882 (95% CI: 0.9879 - 0.9884)

Efficiency (Correct classification rate) = 0.9834 (95% CI: 0.9831 - 0.9837)

Quality index = 0.7161 (95% CI: 0.7112 - 0.7210)

Youden's index = 0.7990 (95% CI: 0.7935 - 0.8046)

Predictive value of positive test = 0.6548 (95% CI: 0.6487 - 0.6609)

Predictive value of negative test = 0.9947 (95% CI: 0.9946 - 0.9949)

False positive rate = 0.0118 (95% CI: 0.0116 - 0.0121)

False negative rate = 0.1891 (95% CI: 0.1836 - 0.1948)

Misclassification rate = 0.0166 (95% CI: 0.0163 - 0.0105)

Prevalence = 0.0270 (95% CI: 0.0266 - 0.0273)

Test level = 0.0334 (95% CI: 0.0330 - 0.0338)

Likelihood ratio of positive test = 68.4853 (95% CI: 66.9522 - 70.0534)

Likelihood ratio of negative test = 5.2245 (95% CI: 5.0733 - 5.3803)

Odds ratio = 357.8040 (95% CI: 342.9821 - 373.2663)

Odds ratio (Haldane's estimator) = 357.7443 (95% CI: 342.9257 - 373.2032)

Cohen's Kappa = 0.7161 (95% CI: 0.7112 - 0.7210)

This kappa indicates substantial agreement.

Test of Ho: Kappa=0: z=606.05, p=0.0000 L.L.L.

Observed Agreement = 0.9834 (95% CI: 0.9831 - 0.9837)

Chance Agreement = 0.9415 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.7245 (95% CI: 0.7198 - 0.7293)

Negative Agreement = 0.9914 (95% CI: 0.9913 - 0.9916)

Dice's Index = 0.7245 (95% CI: 0.7198 - 0.7293)

Yule's Q (Gamma) = 0.9944 (95% CI: 0.9942 - 0.9947)

Pearson Correlation (Phi) = 0.7204 (95% CI: 0.7156 - 0.7252)

Scott's agreement index = 0.7160

Tetrachoric Correlation = 0.9876 (95% CI: 0.9872 - 0.9880)

Goodman &amp; Kruskal's tau (Crit. dep.) = 0.5190 (95% CI: 0.5121 - 0.5259)

Lambda(Symmetric) = 0.4491 (95% CI: 0.4395 - 0.4587)

Lambda(Criterion dep.) = 0.3835 (95% CI: 0.3711 - 0.3959)

Uncertainty Coefficient (Symmetric) = 0.5229 (95% CI: 0.5169 - 0.5288)

Uncertainty Coeff. (Criterion dep.) = 0.5699 (95% CI: 0.5636 - 0.5761)

Pearson Chi Square = 367297.0485, df=1, p=0.000000

Pearson Chi Square with Yates's correction = 367272.2839, df=1, p=0.000000

Likelihood Ratio Chi Square = 100010.8292, df=1, p=0.000000

McNemar's Test = 1756.4004, df=1, p=0.000000

McNemar's Test with Yates's correction = 1755.6276, df=1, p=0.000000



# Appendix 4

**Table 4A - Diagnostic and agreement statistics for UGIB (Model M1).**

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)		Test			
		Positive	Negative		
		3268	2012	5280	
	Positive	5631	697005	702636	
	Negative	8899	699017	707916	
Confidence Intervals					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	61.89%	0.87%	60.57%	63.21%
Sensitivity of a random test	SE_RAN	1.26%			
Quality index of sensitivity	$\kappa(1,0)$	61.41%			
Specificity	SP	99.20%	0.01%	99.18%	99.22%
Specificity of a random test	SP_RAN	98.74%			
Quality index of specificity	$\kappa(0,0)$	36.25%			
Efficiency (Correct classification rate)	EFF	98.92%	0.01%	98.90%	98.94%
Efficiency of a random test	EFF_RAN	98.02%			
Quality index	$\kappa(5,0)$	45.59%	0.52%	44.58%	46.61%
Youden's index	J	61.09%	0.87%	59.78%	62.40%
Predictive value of positive test	PVP	36.72%	0.51%	35.72%	37.73%
Pred. value of a positive random test	PVP_RAN	0.75%			
Predictive value of negative test	PVN	99.71%	0.01%	99.70%	99.72%
Pred. value of a negative random test	PVN_RAN	99.25%			
False positive rate	FP	0.80%	0.01%	0.78%	0.82%
False negative rate	FN	38.11%	0.87%	36.79%	39.43%
Misclassification rate	1-EFF	1.08%	0.01%	1.06%	0.59%
Prevalence	P	0.75%	0.01%	0.73%	0.77%
Test level	Q	1.26%	0.01%	1.23%	1.28%
Likelihood ratio of positive test	LR+	77.2312	1.0173	74.6842	79.8652
Likelihood ratio of negative test	LR-	2.6032	1.0177	2.5153	2.6943
Odds ratio	OR	201.0502	1.0318	189.0732	213.7858
Odds ratio (Haldane's estimator)	OR'	201.0133	1.0318	189.0396	213.7454
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.4559	0.0052	0.4456	0.4661
This kappa indicates moderate agreement.					
Test of Ho: Kappa=0: z=396.96, p=0.0000 LLL					
Observed Agreement	PO	0.9892	0.0001	0.9890	0.9894
Chance Agreement	PE	0.9802			
Positive Agreement	PA	0.4610	0.0052	0.4508	0.4711
Negative Agreement	NA	0.9945	0.0001	0.9944	0.9947
Byrt's Bias Index	BI	-0.0051			
Byrt's Prevalence Asymmetry Index	PI	0.9800			
Bias Adjusted Kappa	BAK	0.4555			
Prevalence & Bias Adjusted Kappa	PABAK	0.9784			
Alternative Indices of Association					
Dice's Index	p(s)	0.4610	0.0052	0.4508	0.4711
Yule's Q (Gamma)	Y	0.9901	0.0003	0.9895	0.9907
Pearson Correlation (Phi)	$\phi$	0.4718	0.0051	0.4617	0.4819
Scott's agreement Index	$\pi$	0.4555			
Tetrachoric Correlation	r(t)	0.9787	0.0005	0.9777	0.9796
Goodman & Kruskal's tau (Crit. dep.)	$\tau(\text{any})$	0.2226	0.0048	0.2131	0.2321
Lambda(Symmetric)	$\lambda(\text{sym})$	0.0886	0.0048	0.0792	0.0980
Lambda(Criterion dep.)	$\lambda(\text{asy})$	0.0000	0.0000	0.0000	0.0000
Uncertainty Coefficient (Symmetric)	U(sym)	0.2913	0.0046	0.2822	0.3004
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.3693	0.0058	0.3582	0.3803
Pearson Chi Square					
Pearson Chi Square	Test	Chi Square	P		
with Yate's correction		157577.17	0.0000		
Likelihood Ratio Chi Square		22985.23	0.0000		
Minimum Expected Frequency					
Minimum Expected Frequency		68.37			
Cells with Expected Frequency < 5		0 of 4 (0%)			
Cells with Expected Frequency < 1		0 of 4 (0%)			
McNemar's Test					
McNemar's Test	Test	Chi Square	P		
with Yate's correction		1713.62	0.0000		
		1712.67	0.0000		
Kullback-Leibler Distance					
D(f  g)		3.05499			
D(g  f)		-0.98392			
Pin		21.22093			
Pout		0.37384			

		Test		
		Positive	Negative	
		Row 61.89% Col 36.72%	Row 38.11% Col 0.29%	0.75%
	Positive	Row 0.80% Col 63.28%	Row 99.20% Col 99.71%	99.25%
	Negative	1.26%	98.74%	N = 707916

No. Decimal Places

Summary

Sensitivity = 0.6189 (95% CI: 0.6057 - 0.6321)

Specificity = 0.9920 (95% CI: 0.9918 - 0.9922)

Efficiency (Correct classification rate) = 0.9892 (95% CI: 0.9890 - 0.9894)

Quality index = 0.4559 (95% CI: 0.4456 - 0.4661)

Youden's index = 0.6109 (95% CI: 0.5978 - 0.6240)

Predictive value of positive test = 0.3672 (95% CI: 0.3572 - 0.3773)

Predictive value of negative test = 0.9971 (95% CI: 0.9970 - 0.9972)

False positive rate = 0.0080 (95% CI: 0.0078 - 0.0082)

False negative rate = 0.3811 (95% CI: 0.3679 - 0.3943)

Misclassification rate = 0.0108 (95% CI: 0.0106 - 0.0059)

Prevalence = 0.0075 (95% CI: 0.0073 - 0.0077)

Test level = 0.0126 (95% CI: 0.0123 - 0.0128)

Likelihood ratio of positive test = 77.2312 (95% CI: 74.6842 - 79.8652)

Likelihood ratio of negative test = 2.6032 (95% CI: 2.5153 - 2.6943)

Odds ratio = 201.0502 (95% CI: 189.0732 - 213.7858)

Odds ratio (Haldane's estimator) = 201.0133 (95% CI: 189.0396 - 213.7454)

Cohen's Kappa = 0.4559 (95% CI: 0.4456 - 0.4661)

This kappa indicates moderate agreement.

Test of Ho: Kappa=0: z=396.96, p=0.0000 LLL

Observed Agreement = 0.9892 (95% CI: 0.9890 - 0.9894)

Chance Agreement = 0.9802 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.4610 (95% CI: 0.4508 - 0.4711)

Negative Agreement = 0.9945 (95% CI: 0.9944 - 0.9947)

Dice's Index = 0.4610 (95% CI: 0.4508 - 0.4711)

Yule's Q (Gamma) = 0.9901 (95% CI: 0.9895 - 0.9907)

Pearson Correlation (Phi) = 0.4718 (95% CI: 0.4617 - 0.4819)

Scott's agreement Index = 0.4555

Tetrachoric Correlation = 0.9787 (95% CI: 0.9777 - 0.9796)

Goodman &amp; Kruskal's tau (Crit. dep.) = 0.2226 (95% CI: 0.2131 - 0.2321)

Lambda(Symmetric) = 0.0886 (95% CI: 0.0792 - 0.0980)

Lambda(Criterion dep.) = 0.0000 (95% CI: 0.0000 - 0.0000)

Uncertainty Coefficient (Symmetric) = 0.2913 (95% CI: 0.2822 - 0.3004)

Uncertainty Coeff. (Criterion dep.) = 0.3693 (95% CI: 0.3582 - 0.3803)

Pearson Chi Square = 157577.1737, df=1, p=0.000000

Pearson Chi Square with Yate's correction = 157527.9596, df=1, p=0.000000

Likelihood Ratio Chi Square = 22985.2348, df=1, p=0.000000

McNemar's Test = 1713.6152, df=1, p=0.000000

McNemar's Test with Yate's correction = 1712.6683, df=1, p=0.000000

Table 4B - Diagnostic and agreement statistics for UGIB (Model M2).

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)		Test			
		Positive	Negative		
		3890	1390	5280	
		5144	697492	702636	
		9034	698882	707916	
Confidence Intervals		95%			
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	73.67%	0.61%	72.46%	74.86%
Sensitivity of a random test	SE_RAN	1.28%			
Quality index of sensitivity	$\kappa(1,0)$	73.33%			
Specificity	SP	99.27%	0.01%	99.25%	99.29%
Specificity of a random test	SP_RAN	98.72%			
Quality index of specificity	$\kappa(0,0)$	42.63%			
Efficiency (Correct classification rate)	EFF	99.08%	0.01%	99.05%	99.10%
Efficiency of a random test	EFF_RAN	98.00%			
Quality Index	$\kappa(5,0)$	53.92%	0.51%	52.93%	54.91%
Youden's Index	J	72.94%	0.61%	71.75%	74.13%
Predictive value of positive test	PVP	43.06%	0.52%	42.04%	44.06%
Pred. value of a positive random test	PVP_RAN	0.75%			
Predictive value of negative test	PVN	99.80%	0.01%	99.79%	99.81%
Pred. value of a negative random test	PVN_RAN	99.25%			
False positive rate	FP	0.73%	0.01%	0.71%	0.75%
False negative rate	FN	26.33%	0.61%	25.14%	27.54%
Misclassification rate	1-EFF	0.92%	0.01%	0.90%	0.41%
Prevalence	P	0.75%	0.01%	0.73%	0.77%
Test level	Q	1.28%	0.01%	1.25%	1.30%
Likelihood ratio of positive test	LR+	100.6341	1.0163	97.4996	103.8694
Likelihood ratio of negative test	LR-	3.7708	1.0233	3.6044	3.9448
Odds ratio	OR	379.4662	1.0348	354.8367	405.8052
Odds ratio (Haldane's estimator)	OR'	379.3419	1.0348	354.7236	405.6687
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.5392	0.0051	0.5293	0.5491
This kappa indicates moderate agreement. Test of Ho: Kappa=0: z=470.45, p =0.0000 L.L.L.					
Observed Agreement	PO	0.9908	0.0001	0.9905	0.9910
Chance Agreement	PE	0.9800			
Positive Agreement	PA	0.5435	0.0050	0.5337	0.5534
Negative Agreement	NA	0.9953	0.0001	0.9952	0.9955
Byrt's Bias Index	BI	-0.0053			
Byrt's Prevalence Asymmetry Index	PI	0.9798			
Bias Adjusted Kappa	BAK	0.5389			
Prevalence & Bias Adjusted Kappa	PABAK	0.9815			
Alternative Indices of Association					
Dice's Index	p(s)	0.5435	0.0050	0.5337	0.5534
Yule's Q (Gamma)	Y	0.9947	0.0002	0.9944	0.9951
Pearson Correlation (Phi)	$\phi$	0.5591	0.0048	0.5497	0.5686
Scott's agreement index	$\pi$	0.5389			
Tetrachoric Correlation	r(t)	0.9883	0.0003	0.9877	0.9888
Goodman & Kruskal's tau (Crit. dep.)	$\tau(\text{asy})$	0.3126	0.0054	0.3020	0.3232
Lambda(Symmetric)	$\lambda(\text{sym})$	0.1747	0.0043	0.1663	0.1830
Lambda(Criterion dep.)	$\lambda(\text{asy})$	0.0000	0.0000	0.0000	0.0000
Uncertainty Coefficient (Symmetric)	U(sym)	0.3752	0.0049	0.3656	0.3648
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.4792	0.0057	0.4680	0.4904
Pearson Chi Square					
Pearson Chi Square	Test	Chi Square	P		
with Yates's correction		221319.22	0.0000		
Likelihood Ratio Chi Square		221261.33	0.0000		
		29829.14	0.0000		
Minimum Expected Frequency					
Minimum Expected Frequency		67.38			
Cells with Expected Frequency < 5		0 of 4 (0%)			
Cells with Expected Frequency < 1		0 of 4 (0%)			
McNemar's Test					
McNemar's Test	Test	Chi Square	P		
with Yates's correction		2156.80	0.0000		
		2155.65	0.0000		
Kullback-Leibler Distance					
D(f  g)		3.74690			
D(g  f)		-1.35132			
Pin		42.38931			
Pout		0.25890			

		Test		
		Positive	Negative	
		Row 73.67% Col 43.06%	Row 26.33% Col 0.20%	0.75%
Positive	Row 0.73% Col 54.94%	Row 99.27% Col 99.80%	99.25%	
Negative	1.28%	98.72%	N = 707916	

No. Decimal Places

Sensitivity = 0.7367 (95% CI: 0.7246 - 0.7486)

Specificity = 0.9927 (95% CI: 0.9925 - 0.9929)

Efficiency (Correct classification rate) = 0.9908 (95% CI: 0.9905 - 0.9910)

Quality index = 0.5392 (95% CI: 0.5293 - 0.5491)

Youden's index = 0.7294 (95% CI: 0.7175 - 0.7413)

Predictive value of positive test = 0.4306 (95% CI: 0.4204 - 0.4409)

Predictive value of negative test = 0.9980 (95% CI: 0.9979 - 0.9981)

False positive rate = 0.0073 (95% CI: 0.0071 - 0.0075)

False negative rate = 0.2633 (95% CI: 0.2514 - 0.2754)

Misclassification rate = 0.0092 (95% CI: 0.0090 - 0.0094)

Prevalence = 0.0075 (95% CI: 0.0073 - 0.0077)

Test level = 0.0128 (95% CI: 0.0125 - 0.0130)

Likelihood ratio of positive test = 100.6341 (95% CI: 97.4996 - 103.8694)

Likelihood ratio of negative test = 3.7708 (95% CI: 3.6044 - 3.9448)

Odds ratio = 379.4662 (95% CI: 354.8367 - 405.8052)

Odds ratio (Haldane's estimator) = 379.3419 (95% CI: 354.7236 - 405.6687)

Cohen's Kappa = 0.5392 (95% CI: 0.5293 - 0.5491)

This kappa indicates moderate agreement.

Test of Ho: Kappa=0: z=470.45, p=0.0000 LLL

Observed Agreement = 0.9908 (95% CI: 0.9905 - 0.9910)

Chance Agreement = 0.9800 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.5435 (95% CI: 0.5337 - 0.5534)

Negative Agreement = 0.9953 (95% CI: 0.9952 - 0.9955)

Dice's Index = 0.5435 (95% CI: 0.5337 - 0.5534)

Yule's Q (Gamma) = 0.9947 (95% CI: 0.9944 - 0.9951)

Pearson Correlation (Phi) = 0.5591 (95% CI: 0.5497 - 0.5686)

Scott's agreement index = 0.5389

Tetrachoric Correlation = 0.9883 (95% CI: 0.9877 - 0.9888)

Goodman & Kruskal's tau (Crit. dep.) = 0.3126 (95% CI: 0.3020 - 0.3232)

Lambda(Symmetric) = 0.1747 (95% CI: 0.1663 - 0.1830)

Lambda(Criterion dep.) = 0.0000 (95% CI: 0.0000 - 0.0000)

Uncertainty Coefficient (Symmetric) = 0.3752 (95% CI: 0.3656 - 0.3848)

Uncertainty Coeff. (Criterion dep.) = 0.4792 (95% CI: 0.4680 - 0.4904)

Pearson Chi Square = 221319.2214, df=1, p=0.000000

Pearson Chi Square with Yates's correction = 221261.3279, df=1, p=0.000000

Likelihood Ratio Chi Square = 29829.1354, df=1, p=0.000000

McNemar's Test = 2156.7977, df=1, p=0.000000

McNemar's Test with Yates's correction = 2155.6488, df=1, p=0.000000

Table 4C - Diagnostic and agreement statistics for UGIB (Model M3).

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)	Positive  Negative	Test		5280  702636	70.90%
		Positive	Negative		
		3678	1602		
		4144	698492		
		7822	700094	707916	
Confidence Intervals					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	89.66%	0.63%	68.40%	70.90%
Sensitivity of a random test	SE_RAN	1.10%			
Quality index of sensitivity	$\kappa(1,0)$	89.32%			
Specificity	SP	99.41%	0.01%	99.39%	99.43%
Specificity of a random test	SP_RAN	98.90%			
Quality Index of specificity	$\kappa(0,0)$	48.62%			
Efficiency (Correct classification rate)	EFF	99.19%	0.01%	99.17%	99.21%
Efficiency of a random test	EFF_RAN	98.17%			
Quality index	$\kappa(1,0)$	55.75%	0.52%	54.73%	56.77%
Youden's Index	J	89.07%	0.63%	67.83%	70.31%
Predictive value of positive test	PVP	47.02%	0.56%	45.91%	48.13%
Pred. value of a positive random test	PVP_RAN	0.75%			
Predictive value of negative test	PVN	99.77%	0.01%	99.76%	99.78%
Pred. value of a negative random test	PVN_RAN	99.25%			
False positive rate	FP	0.59%	0.01%	0.57%	0.61%
False negative rate	FN	30.34%	0.63%	29.10%	31.60%
Misclassification rate	1-EFF	0.81%	0.01%	0.79%	0.47%
Prevalence	P	0.75%	0.01%	0.73%	0.77%
Test level	Q	1.10%	0.01%	1.08%	1.13%
Likelihood ratio of positive test	LR+	118.1105	1.0181	114.0283	122.3409
Likelihood ratio of negative test	LR-	3.2764	1.0211	3.1452	3.4131
Odds ratio	OR	386.9821	1.0343	362.2143	413.4436
Odds ratio (Haldane's estimator)	OR*	386.8676	1.0343	362.1098	413.3181
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.5575	0.0052	0.5473	0.5677
This kappa indicates moderate agreement. Test of Ho: Kappa=0: z=478.32, p=0.0000 LLL					
Observed Agreement	PO	0.9919	0.0001	0.9917	0.9921
Chance Agreement	PE	0.9817			
Positive Agreement	PA	0.5614	0.0052	0.5513	0.5716
Negative Agreement	NA	0.9959	0.0001	0.9958	0.9960
Byrt's Bias Index	BI	-0.0036			
Byrt's Prevalence Asymmetry Index	PI	0.9815			
Bias Adjusted Kappa	BAK	0.5573			
Prevalence & Bias Adjusted Kappa	PABAK	0.9638			
Alternative Indices of Association					
Dice's Index	p(s)	0.5614	0.0052	0.5513	0.5716
Yule's Q (Gamma)	$\gamma$	0.9948	0.0002	0.9945	0.9952
Pearson Correlation (Phi)	$\phi$	0.5685	0.0051	0.5585	0.5785
Scott's agreement index	$\pi$	0.5573			
Tetrachoric Correlation	r(t)	0.9885	0.0003	0.9879	0.9890
Goodman & Kruskal's tau (Crit. dep.)	$\tau(asy)$	0.3232	0.0058	0.3118	0.3346
Lambda(Symmetric)	$\lambda(sym)$	0.1584	0.0048	0.1490	0.1679
Lambda(Criterion dep.)	$\lambda(asy)$	0.0000	0.0000	0.0000	0.0000
Uncertainty Coefficient (Symmetric)	U(sym)	0.3878	0.0052	0.3777	0.3979
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.4619	0.0059	0.4503	0.4734
Test					
Test	Chi Square	P			
Pearson Chi Square	228792.69	0.0000			
with Yates's correction	228729.48	0.0000			
Likelihood Ratio Chi Square	28751.27	0.0000			
Minimum Expected Frequency = 58.34					
Cells with Expected Frequency < 5 = 0 of 4 (0%)					
Cells with Expected Frequency < 1 = 0 of 4 (0%)					
Test					
Test	Chi Square	P			
McNemar's Test	1124.57	0.0000			
with Yates's correction	1123.68	0.0000			
Kullback-Leibler Distance					
D(f  g)	3.68394				
D(g  f)	-1.20790				
Pin	39.80293				
Pout	0.29882				

	Test		
	Positive	Negative	
	Row 99.66% Col 47.02%	Row 30.34% Col 0.23%	0.75%
Positive	Row 0.59% Col 52.98%	Row 99.41% Col 99.77%	99.25%
	1.10%	98.90%	N = 707916

No. Decimal Places

4

## Summary

Sensitivity = 0.6966 (95% CI: 0.6840 - 0.7090)

Specificity = 0.9941 (95% CI: 0.9939 - 0.9943)

Efficiency (Correct classification rate) = 0.9919 (95% CI: 0.9917 - 0.9921)

Quality index = 0.5575 (95% CI: 0.5473 - 0.5677)

Youden's index = 0.6907 (95% CI: 0.6783 - 0.7031)

Predictive value of positive test = 0.4702 (95% CI: 0.4591 - 0.4813)

Predictive value of negative test = 0.9977 (95% CI: 0.9976 - 0.9978)

False positive rate = 0.0059 (95% CI: 0.0057 - 0.0061)

False negative rate = 0.3034 (95% CI: 0.2910 - 0.3160)

Misclassification rate = 0.0081 (95% CI: 0.0079 - 0.0083)

Prevalence = 0.0075 (95% CI: 0.0073 - 0.0077)

Test level = 0.0110 (95% CI: 0.0108 - 0.0113)

Likelihood ratio of positive test = 118.1105 (95% CI: 114.0283 - 122.3409)

Likelihood ratio of negative test = 3.2764 (95% CI: 3.1452 - 3.4131)

Odds ratio = 386.9821 (95% CI: 362.2143 - 413.4436)

Odds ratio (Haldane's estimator) = 386.8676 (95% CI: 362.1098 - 413.3181)

Cohen's Kappa = 0.5575 (95% CI: 0.5473 - 0.5677)

This kappa indicates moderate agreement.

Test of Ho: Kappa=0: z=478.32, p=0.0000 LLL

Observed Agreement = 0.9919 (95% CI: 0.9917 - 0.9921)

Chance Agreement = 0.9817 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.5614 (95% CI: 0.5513 - 0.5716)

Negative Agreement = 0.9959 (95% CI: 0.9958 - 0.9960)

Dice's Index = 0.5614 (95% CI: 0.5513 - 0.5716)

Yule's Q (Gamma) = 0.9948 (95% CI: 0.9945 - 0.9952)

Pearson Correlation (Phi) = 0.5685 (95% CI: 0.5585 - 0.5785)

Scott's agreement index = 0.5573

Tetrachoric Correlation = 0.9885 (95% CI: 0.9879 - 0.9890)

Goodman &amp; Kruskal's tau (Crit. dep.) = 0.3232 (95% CI: 0.3118 - 0.3346)

Lambda(Symmetric) = 0.1584 (95% CI: 0.1490 - 0.1679)

Lambda(Criterion dep.) = 0.0000 (95% CI: 0.0000 - 0.0000)

Uncertainty Coefficient (Symmetric) = 0.3878 (95% CI: 0.3777 - 0.3979)

Uncertainty Coeff. (Criterion dep.) = 0.4619 (95% CI: 0.4503 - 0.4734)

Pearson Chi Square = 228792.6871, df=1, p=0.000000

Pearson Chi Square with Yates's correction = 228729.4831, df=1, p=0.000000

Likelihood Ratio Chi Square = 28751.2731, df=1, p=0.000000

McNemar's Test = 1124.5674, df=1, p=0.000000

McNemar's Test with Yates's correction = 1123.6827, df=1, p=0.000000

Table 4D - Diagnostic and agreement statistics for UGIB (Model M4).

Diagnostic & Agreement Statistics					
Criterion (Gold Standard)		Test		5277	702636
		Positive	Negative		
	Positive	3434	1843		
	Negative	1701	700935		
		5135	702778	707913	
Confidence Intervals					
Index	Symbol	Estimate	se	Lower 95% CI	Upper 95% CI
Sensitivity	SE	65.07%	0.66%	63.77%	66.36%
Sensitivity of a random test	SE_RAN	0.73%			
Quality index of sensitivity	$\kappa(1,0)$	84.82%			
Specificity	SP	99.78%	0.01%	99.75%	99.77%
Specificity of a random test	SP_RAN	99.27%			
Quality index of specificity	$\kappa(0,0)$	86.63%			
Efficiency (Correct classification rate)	EFF	99.50%	0.01%	99.48%	99.52%
Efficiency of a random test	EFF_RAN	98.54%			
Quality index	$\kappa(1,0)$	65.71%	0.54%	64.65%	66.77%
Youden's index	J	64.83%	0.86%	63.55%	66.12%
Predictive value of positive test	PVP	66.87%	0.66%	65.57%	68.16%
Pred. value of a positive random test	PVP_RAN	0.75%			
Predictive value of negative test	PVN	99.74%	0.01%	99.73%	99.75%
Pred. value of a negative random test	PVN_RAN	99.25%			
False positive rate	FP	0.24%	0.01%	0.23%	0.25%
False negative rate	FN	34.93%	0.66%	33.64%	36.23%
Misclassification rate	1-EFF	0.50%	0.01%	0.48%	0.54%
Prevalence	P	0.75%	0.01%	0.73%	0.77%
Test level	Q	0.73%	0.01%	0.71%	0.75%
Likelihood ratio of positive test	LR+	268.8062	1.0266	255.3346	282.9885
Likelihood ratio of negative test	LR-	2.8563	1.0190	2.7531	2.9635
Odds ratio	OR	767.8005	1.0384	713.0788	826.7216
Odds ratio (Haldane's estimator)	OR*	767.4790	1.0384	712.7869	826.3676
Kappa and Related Indices					
Cohen's Kappa	$\kappa$	0.6571	0.0054	0.6465	0.6677
This kappa indicates substantial agreement.					
Test of Ho: Kappa=0: z=552.92, p=0.0000 LLL					
Observed Agreement	PO	0.9950	0.0001	0.9948	0.9952
Chance Agreement	PE	0.9854			
Positive Agreement	PA	0.6596	0.0054	0.6491	0.6702
Negative Agreement	NA	0.9975	0.0000	0.9974	0.9976
Byrt's Bias Index	BI	0.0002			
Byrt's Prevalence Asymmetry Index	PI	0.9853			
Bias Adjusted Kappa	BAK	0.6571			
Prevalence & Bias Adjusted Kappa	PABAK	0.9900			
Alternative Indices of Association					
Dice's Index	p(s)	0.6596	0.0054	0.6491	0.6702
Yule's Q (Gamma)	y	0.9974	0.0001	0.9972	0.9976
Pearson Correlation (Phi)	$\phi$	0.6572	0.0054	0.6466	0.6678
Scott's agreement index	$\pi$	0.6571			
Tetrachoric Correlation	r(t)	0.9940	0.0002	0.9937	0.9943
Goodman & Kruskal's tau (Crit. dep.)	$\tau(asy)$	0.4319	0.0071	0.4179	0.4458
Lambda(Symmetric)	$\lambda(sym)$	0.3182	0.0108	0.2982	0.3403
Lambda(Criterion dep.)	$\lambda(cry)$	0.3284	0.0111	0.3066	0.3502
Uncertainty Coefficient (Symmetric)	U(sym)	0.4894	0.0059	0.4777	0.5010
Uncertainty Coeff. (Criterion dep.)	U(asy)	0.4839	0.0064	0.4714	0.4963
Chi Square					
Pearson Chi Square		305722.82	0.0000		
with Yate's correction		305632.80	0.0000		
Likelihood Ratio Chi Square		30106.67	0.0000		
Minimum Expected Frequency		38.28			
Cells with Expected Frequency < 5		0 of 4 (0%)			
Cells with Expected Frequency < 1		0 of 4 (0%)			
McNemar's Test					
McNemar's Test		5.89	0.0171		
with Yate's correction		5.61	0.0179		
Kullback-Leibler Distance					
D(f  g)		4.00683			
D(g  f)		-1.06054			
Pin		54.97257			
Pout		0.34627			

	Test		0.75%
	Positive	Negative	
	Row 61.07% Col 66.87%	Row 34.93% Col 0.26%	
	Row 0.24% Col 31.13%	Row 99.76% Col 99.74%	
		0.73%	99.25%
		0.73%	99.27%
		N = 707913	

No. Decimal Places

Summary  
Sensitivity = 0.6507 (95% CI: 0.6377 - 0.6636)

Specificity = 0.9976 (95% CI: 0.9975 - 0.9977)

Efficiency (Correct classification rate) = 0.9950 (95% CI: 0.9948 - 0.9952)

Quality index = 0.6571 (95% CI: 0.6465 - 0.6677)

Youden's index = 0.6483 (95% CI: 0.6355 - 0.6612)

Predictive value of positive test = 0.6687 (95% CI: 0.6557 - 0.6816)

Predictive value of negative test = 0.9974 (95% CI: 0.9973 - 0.9975)

False positive rate = 0.0024 (95% CI: 0.0023 - 0.0025)

False negative rate = 0.3493 (95% CI: 0.3364 - 0.3623)

Misclassification rate = 0.0050 (95% CI: 0.0048 - 0.0054)

Prevalence = 0.0075 (95% CI: 0.0073 - 0.0077)

Test level = 0.0073 (95% CI: 0.0071 - 0.0075)

Likelihood ratio of positive test = 268.8062 (95% CI: 255.3346 - 282.9885)

Likelihood ratio of negative test = 2.8563 (95% CI: 2.7531 - 2.9635)

Odds ratio = 767.8005 (95% CI: 713.0788 - 826.7216)

Odds ratio (Haldane's estimator) = 767.4790 (95% CI: 712.7869 - 826.3676)

Cohen's Kappa = 0.6571 (95% CI: 0.6465 - 0.6677)

This kappa indicates substantial agreement.

Test of Ho: Kappa=0: z=552.92, p=0.0000 LLL

Observed Agreement = 0.9950 (95% CI: 0.9948 - 0.9952)

Chance Agreement = 0.9854 (95% CI: 0.0000 - 0.0000)

Positive Agreement = 0.6596 (95% CI: 0.6491 - 0.6702)

Negative Agreement = 0.9975 (95% CI: 0.9974 - 0.9976)

Dice's Index = 0.6596 (95% CI: 0.6491 - 0.6702)

Yule's Q (Gamma) = 0.9974 (95% CI: 0.9972 - 0.9976)

Pearson Correlation (Phi) = 0.6572 (95% CI: 0.6466 - 0.6678)

Scott's agreement index = 0.6571

Tetrachoric Correlation = 0.9940 (95% CI: 0.9937 - 0.9943)

Goodman &amp; Kruskal's tau (Crit. dep.) = 0.4319 (95% CI: 0.4179 - 0.4458)

Lambda(Symmetric) = 0.3182 (95% CI: 0.2982 - 0.3403)

Lambda(Criterion dep.) = 0.3284 (95% CI: 0.3066 - 0.3502)

Uncertainty Coefficient (Symmetric) = 0.4894 (95% CI: 0.4777 - 0.5010)

Uncertainty Coeff. (Criterion dep.) = 0.4839 (95% CI: 0.4714 - 0.4963)

Pearson Chi Square = 305722.8210, df=1, p=0.000000

Pearson Chi Square with Yate's correction = 305632.7959, df=1, p=0.000000

Likelihood Ratio Chi Square = 30106.6691, df=1, p=0.000000

McNemar's Test = 5.6896, df=1, p=0.017066

McNemar's Test with Yate's correction = 5.6098, df=1, p=0.017861

